5-2013

DEVELOPMENT OF NOVEL METHODS TO MINIMIZE THE IMPACT OF SEQUENCING ERRORS IN THE NEXT-GENERATION SEQUENCING DATA ANALYSIS

Xiaofeng Zheng

DEVELOPMENT OF NOVEL METHODS TO MINIMIZE THE IMPACT OF SEQUENCING ERRORS IN THE NEXT-GENERATION SEQUENCING DATA ANALYSIS

By

Xiaofeng Zheng, M.S.

APPROVED:

_____

Supervisory Professor, Shoudan Liang, PhD

_____

Peter Müller, PhD

_____

Yuan Ji, PhD

_____

Qing Ma, PhD

_____

Marcos R. Estecio, PhD

APPROVED:

_____

Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

DEVELOPMENT OF NOVEL METHODS TO MINIMIZE THE IMPACT OF SEQUENCING ERRORS IN THE NEXT-GENERATION SEQUENCING DATA ANALYSIS

A

DISSERTATION

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
and

The University of Texas
MD Anderson Cancer Center
Graduate School of Biomedical Sciences
in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

By

Xiaofeng Zheng, M.S.

Houston, Texas,

May, 2013

To my parents, Zhenbo and Yahua, for their unconditional love, inspiration and sacrifice;

My husband, Ying, without whom I could never have reached this far;

My beloved son, Nolan.

I love you all!

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Shoudan Liang for taking me under his wing as his student and sharing with me his passion and knowledge on bioinformatics research. I attribute my Ph.D. degree to his encouragement and dedication.

I would also like to thank my committee members: Drs. Peter Müller, Yuan Ji, Qing Ma, Marcos R. Estecio and previous members, for their time and advice with regard to my thesis.

This work would not be possible without the collaborations with Dr. Yuan Ji and Dr. Marcos R. Estecio, Special thanks are owed to them for their wonderful suggestions and kind help with my project.

I extend my thanks to my lab mates, fellow students, faculty and stuff in the Department of Bioinformatics & Computational Biology for the friendship we have built and the many conversations we have had throughout the years while working, playing and traveling together. They have made my life easy and colorful.

Last, but certainly not least, I would like to thank my parents, Zhenbo and Yahua, for their love over the years and for encouraging me to pursue my dreams; my husband, Ying, for his unending love and support on so many levels; my son, Nolan, for the endless joy he brings to me.

Development of novel methods to minimize the impact of sequencing errors in the next-generation sequencing data analysis

Publication No._____

Xiaofeng Zheng, M.S.

Supervisory Professor: Shoudan Liang, Ph.D

Next-generation sequencing (NGS) technology has become a prominent tool in biological and biomedical research. However, NGS data analysis, such as de novo assembly, mapping and variants detection is far from maturity, and the high sequencing error-rate is one of the major problems. .

To minimize the impact of sequencing errors, we developed a highly robust and efficient method, MTM, to correct the errors in NGS reads. We demonstrated the effectiveness of MTM on both single-cell data with highly non-uniform coverage and normal data with uniformly high coverage, reflecting that MTM's performance does not rely on the coverage of the sequencing reads. MTM was also compared with Hammer and Quake, the best methods for correcting non-uniform and uniform data respectively. For non-uniform data, MTM outperformed both Hammer and Quake. For uniform data, MTM showed better performance than Quake and comparable results to Hammer. By making better error correction with MTM, the quality of downstream analysis, such as mapping and SNP detection, was improved.

SNP calling is a major application of NGS technologies. However, the existence of sequencing errors complicates this process, especially for the low coverage (<5×) data. Since many NGS studies are now based on data with low to medium coverage (<20×), on which most existing SNP calling methods perform poorly, we developed a Bayesian-based approach for calling SNPs that is robust to the sequencing depth. We successfully applied this approach to identify the SNPs in prostate cancer cell line PC-3 and colon cancer cell lines RKO and SW48, whose mutation status are unknown. Our method outperforms the existing methods - Varscan and DNAnexus - by identifying more SNPs while maintaining higher dbSNP rates, especially for the low coverage PC-3 data. In summary, we identified 107 potential causal genes for PC-3. For RKO and SW48 cell lines, 701 and 652 potential causal genes were identified respectively, and 297 genes are in common. With the ability of piggybacking on the ChIP-Seq, RNA-Seq and other data with low or uneven coverage, this approach is expected to have a wide range of applications.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CDS: Coding sequence

CE: Capillary electrophoresis

CG69: Complete Genomics database

ChIP-Seq: Chromatin Immunoprecipitation Sequencing

dbSNP: Single Nucleotide Polymorphism Database

emPCR: emulsion PCR

ESP: NHLBI GO Exome Sequencing Project

GA: Genome analyzer

indel: insertion and deletion

ncRNA: non-coding RNA

NGS: Next-Generation Sequencing

SNP: Single-nucleotide polymorphism

SOLiD: Sequencing by Oligo Ligation Detection

Ti/Tv: Transition/Transversion ratio

UTR: untranslated region

CHAPTER ONE

Introduction & Aims

Next-generation sequencing (NGS), as one of the most influential breakthroughs in the biological sciences in the past decades, revolutionized the genomic research. The introduction of NGS technology has changed the way to acquire genetic information from various species to an unprecedented level on speed and cost.

1. Advance in DNA sequencing technologies

DNA sequencing is the process to examine the nucleotide order of a DNA sequence. Deciphering DNA sequence plays an essential role in biological researches. Since early 1990s, the capillary electrophoresis (CE) - base Sanger sequencing [1-3] has dominated the industry of genome analysis for almost two decades and led to many monumental accomplishments, including finishing a "rough draft" of human genome [4]. However, Sanger sequencing is hampered by its inherent limitation on throughput, scalability and speed. Thus, an entirely new technology is required to overcome such limitations. Sanger sequencing was considered as the first-generation sequencing, and the new sequencing technologies are referred as next-generation sequencing. It has been seven years since the advent of NGS, and the increase of its data output has outpaced Moore's law, at a rate of more than doubling each year. In 2007, a single sequencing run produced the maximal 1 GB data, and in 2011, 1TB data could be produced in a single sequencing run, which is ~1000 times of increase. Meanwhile, the dropping speed of the sequencing cost is faster than Moore's law. These days, more than five human genomes can be sequenced in a single run with the cost of less than $5,000, and analyze the data within one week. In comparison, the Sanger sequencing would take ~10 years to produce these data and additional 3 years to finish the analysis, which costs nearly 3

billion US dollars. **Figure 1.1** shows the cost change of sequencing a human-sized genome.

2. Platforms of NGS

NGS is the technology that can sequence millions of DNA sequences in parallel. It includes two types of techniques, which are distinguished with the names of "second-generation" and "third-generation". Second-generation sequencing works by amplifying the DNA templates immobilized on a solid matrix and sequencing them cyclically, while the third-generation sequencing employ single molecule PCR-free and cycle-free protocols. There are three second-generation platforms: 454 sequencing (Roche Applied Science), Solexa sequencing (Illumina Genome Analyzer), and SOLiD sequencing (Applied Biosystems). The third-generation sequencing such as Pacific Biosciences is still not mature and may take may take a few years to rival the second-generation platforms and become the mainstream of the market. Therefore, we will only discuss the second-generation platforms here. The comparison of these three second-generation sequencing platforms is in **Table 1.1**.

2.1. Roche 454 system

Roche 454 technology is the first NGS technology that was released to the market in 2005. Initially in 2005, the read length of 454 was 100-150 bp, and the output per run is 20Mb. In 2008, the upgraded 454 GS FLX system was able to produce 700bp long reads, the accuracy of which is 99.9% after filtering. On average, 0.7 G data was

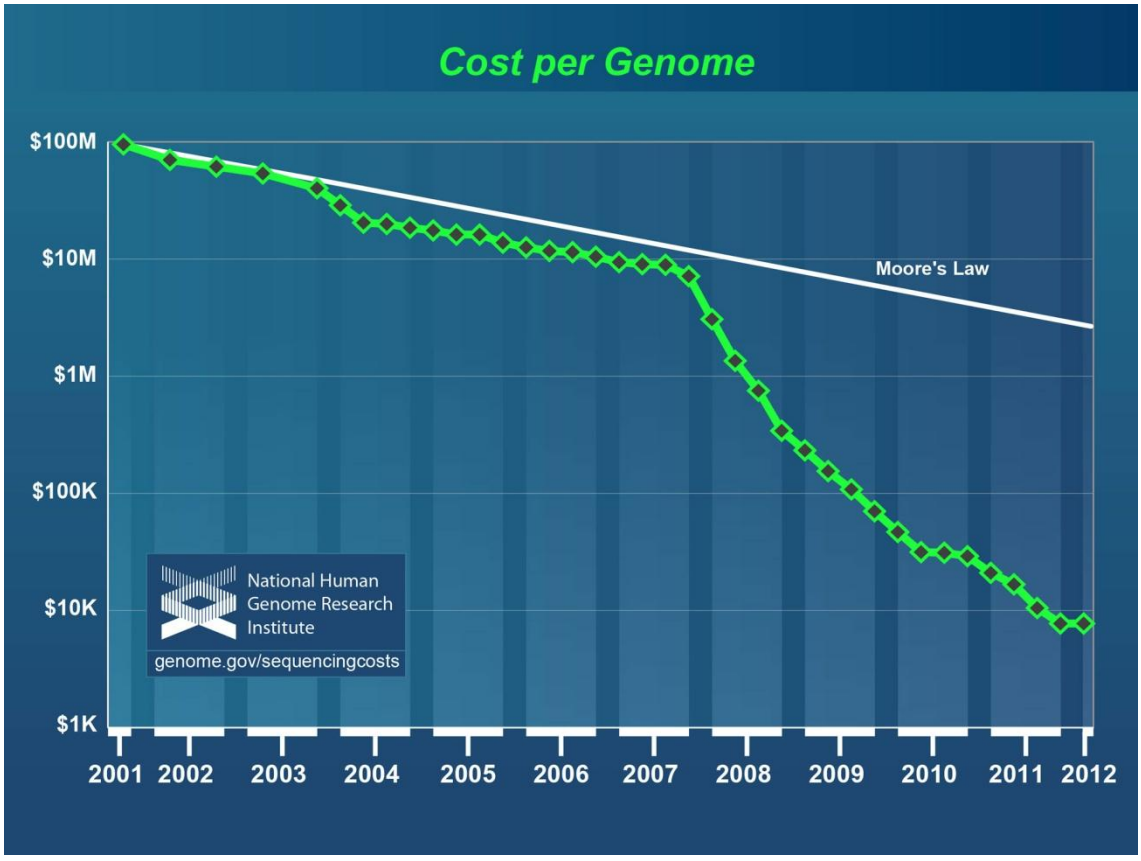


Figure 1.1 Typical cost of sequencing a human-sized genome, on a logarithmic scale. Note that the drastic trend faster than Moore's law beginning in January 2008 as NGS was invented [6].

Table 1.1. Comparison of sequencing platforms [7]. (a) Advantage and mechanism of sequencers. (b) Components and cost of sequencers. (c) Application of sequencers.

(a)

Sequencer	454 GS FLX	HiSeq 2000	SOLiDv4	Sanger 3730xl
Sequencing mechanism	Pyrosequencing	Sequencing by synthesis	Ligation and two-base coding	Dideoxy chain termination
Read length	700 bp	50SE, 50PE, 101PE	50 + 35 bp or 50 + 50 bp	400 ~ 900 bp
Accuracy	99.9%*	98%, (100PE)	99.94% *raw data	99.999%
Reads	1 M	3 G	1200~1400 M	—
Output data/run	0.7 Gb	600 Gb	120 Gb	1.9~84 Kb
Time/run	24 Hours	3~10 Days	7 Days for SE 14 Days for PE	20 Mins~3 Hours
Advantage	Read length, fast	High throughput	Accuracy	High quality, long read length
Disadvantage	Error rate with polybase more than 6, high cost, low throughput	Short read assembly	Short read assembly	High cost low throughput

(b)

Sequencers	454 GS FLX	HiSeq 2000	SOLiDv4	3730xl
Instrument price	Instrument $500,000, $7000 per run	Instrument $690,000, $6000/(30x) human genome	Instrument $495,000, $15,000/100 Gb	Instrument $95,000, about $4 per 800 bp reaction
CPU	2* Intel Xeon X5675	2* Intel Xeon X5560	8* processor 2.0 GHz	Pentium IV 3.0 GHz
Memory	48 GB	48 GB	16 GB	1 GB
Hard disk	1.1 TB	3 TB	10 TB	280 GB
Automation in library preparation	Yes	Yes	Yes	No
Other required device	REM e system	cBot system	EZ beads system	No
Cost/million bases	$10	$0.07	$0.13	$2400

(c)

Sequencers	454 GS FLX	HiSeq 2000	SOLiDv4	3730xl
Resequencing		Yes	Yes	
De novo	Yes	Yes		Yes
Cancer	Yes	Yes	Yes	
Array	Yes	Yes	Yes	Yes
High GC sample	Yes	Yes	Yes	
Bacterial	Yes	Yes	Yes	
Large genome	Yes	Yes		
Mutation detection	Yes	Yes	Yes	Yes

output per run within 24 hours. In late 2009, the output of 454 system had upgraded to 14G per run. The advantages of 454 were its longer read length and faster speed (10 hours from starting to completion). However, the high cost of reagents (about $12.56 per million bases) became a big shortcoming of 454 system. Also, for the poly-bases longer than 6bp, the error rate was relatively high. The comparison of 454 with HiSeq from Illumina and SOLiD are in **Table 1.1**.

2.2. Illumina GA/HiSeq system

Genome Analyzer (GA) was released by Solexa in 2006, and then purchased by Illumina company in 2007. At first, the output of Solexa GA was 1G/run, and then increased to 20G/run in August, 2009, and 30G/run and 50G/run respectively in October and December in the same year. The latest release, GAIIx series, can have 85G/run. In 2010, Illumina adopted the same sequencing strategy with GA to get HiSeq 2000 launched. The output of HiSeq 2000 was 200G/run initially and improved to 600G/run recently and finished in 8 days. In the short run, it is expected to reach 1T/run, and the cost of sequencing a personal genome would drop below 1K US dollars. The average error rate could be below 2% after filtering. As the cheapest platform, HiSeq 2000 is able to sequence one million bases with only$0.02.

2.3. AB SOLiD system

In 2006, after the Solexa was released, SOLiD (Sequencing by Oligo Ligation Detection) entered the market in 2007. Initially, the read length of SOLiD produced 3G 35-bp-long reads per run. By means of the dinucleotide sequencing method, the

accuracy of SOLiD could reach 99.85% after filtering. In late 2010, three years later, with the release of SOLiD 5500xl sequencing system,30G reads with of the length of 85 bp and accuracy of 99.99% were produced in a single run in 7 days. The current cost using SOLiD 5500xl is about $0.04/million bases. But the limitation on de novo sequencing and large genome sequencing is still its major shortcoming.

3. Mechanism for various platforms of NGS

Although the sequencing biochemistry and the array generation are quite diverse for different platforms, their workflows are similar in concept (**Figure 1.2**). Second-generation sequencing follows two principles: DNA templates immobilized and separated on a solid matrix are amplified with DNA polymerase; the replicated DNA are sequenced cyclically. For the library preparation, the DNA sequences are randomly fragmented and ligated to common adaptor sequences in vitro. PCR primers complementary to the adaptor sequences are used to amplify the library immobilized on the support matrix for amplification purpose. Emulsion PCR (emPCR) was employed by both SOLiD and Roche 454 system to clone DNA templates linked to beads [8]. The concentration of beads and template that are added to a water and oil emulsion are controlled carefully to guarantee each emulsion droplet only contains one bead and one DNA template. We call the amplified beads sequencing features. After emPCR, the sequencing features are randomly deposited to pico-wells [11, 12]. Differently from 454 and SOLiD, Solexa employs bridge PCR to generate clonally amplified DNA clusters [13].

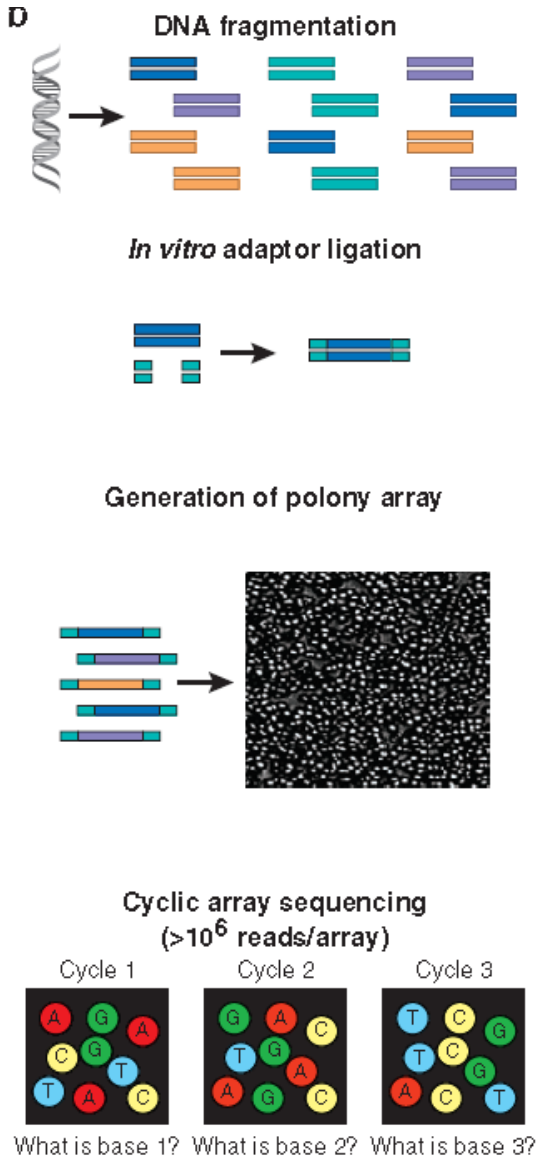


Figure 1.2. Work flow of second-generation sequencing. In shotgun sequencing with cyclic-array methods, common adaptors are ligated to fragmented genomic DNA, which is then subjected to one of several protocols that results in an array of millions of spatially immobilized PCR colonies or "polonies" [9]. Each polony consists of many copies of a single shotgun library fragment. As all polonies are tethered to a planar array, a single microliter-scale reagent volume (e.g., for primer hybridization and then for enzymatic extension reactions) can be applied to manipulate all array features in parallel. Similarly, imaging-based detection of fluorescent labels incorporated with each extension can be used to acquire sequencing data on all features in parallel. Successive iterations of enzymatic interrogation and imaging are used to build up a contiguous sequencing read for each array feature [10].

Each NGS platform uses a unique sequencing chemistries and methods for signal detection. 454 employs pyrosequencing, whereby the nucleotide species is indicated by the chemiluminescent and the number of bases incorporated are correlated to the intensity of signal. Illumina uses reversible dye terminator in each cycle to incorporate a single base, and then image and cleave the terminator in the end. Solid employs sequencing by ligation to measure every base twice by dinucleotide encoding.

Compared with Sanger sequencing, NGS has the following advantages: (1) construction of a sequence library and clonal amplification all in vitro provide the basis of parallel sequencing, which breaks the limitation of conventional sequencing. (2) The replacement of the conventional capillary-based sequencing with array-base sequencing greatly increases the degree of parallelism. As the size the array features are in the scale of micrometer, the imaging process becomes more efficient. (3) The array-based design dramatically reduced the dosage of reagent from the scale of microliters to picoliters or femtoliters per feature drop. All these advantages result in the remarkably lower cost of NGS.

4. Applications of NGS

4.1. Whole genome sequencing

Whole genome sequencing is the sequencing process that determines the complete DNA sequence of organism's genome at a single time. It includes *de novo* sequencing and whole genome resequencing. *De novo* sequencing is the sequencing without prior knowledge of the sequenced genome, and its purpose is to assembly a new

genome with the sequencing reads. De novo sequencing is required for decipher the unknown genomes. As a predominant application of NGS, whole genome resequencing can provide complete genetic information for individual's genome or cancer genome. It is usually used to detect genome-wide single nucleotide variants, indels, copy number variations, and genomic rearrangements [14].

4.2. Targeted Sequencing

Targeted sequencing is the process that only sequence the region that the researchers are interested in. Instead of sequencing the whole genome, this method reduces the time, cost, and providing a higher sequencing coverage. It is usually used to discover the genetic variations by sequencing many individuals. The ability of obtaining high coverage enables NGS to identify rare variants.

Amplicon sequencing is one of the targeted sequencing techniques that sequences selected genome regions of hundreds of base pairs long. Amplicon library can be prepared with commercially available kits, which allow the researchers to prepare the customized targeted region from multiple samples within hours.

Similarly to Amplicon sequencing, target enrichment is also a technique that selectively sequences the genes or regions that researchers are interested in. Differently to Amplicon, target enrichment allows researchers to sequence longer DNA sequences and larger amount of DNA from each sample. A lot of kits are available for the researcher to prepare the library. People can also design their own probes to sequences the regions related to their interest. Exome sequencing, also known as exome capture, is the most popular application of target enrichment approaches. It only sequence the

protein coding regions of the genome. Compared to whole genome sequencing, it is cheaper, but still effective. Exons constitute about only 1% of the human genome [15], which is ~30Mb in length, but about 85% of the disease-causing mutations are associated with these regions [16]. The work flow of exome capture is shown in **Figure 1.3.**

4.3. RNA-Seq

RNA-Seq, also known as an "Whole Transcriptome Sequencing" [17], is an approach that sequences cDNA with NGS technologies to obtain information about the RNA content of a sample. It has been adopted to the disease associated studies and dubbed "a revolutionary tool for transcriptomics" [18]. With the deep coverage and base-level resolution, RNA-Seq is primarily used to study the gene expression profiling, including gene alleles and differently spliced transcripts [19]. Meanwhile, RNA-Seq is often used to provide information about non-coding RNAs, post-transcriptional variants, and gene fusions. However, the significantly different expression levels between genes usually result in insufficient coverage to accurately call variants.

The RNA library preparation varies for different platforms of NGS [18], each of which has several kits designed to build different types of libraries. However, the workflows of different sequencing technologies are similar in concept. To separate the coding RNA from non-coding RNA, poly(T) oligos are designed to covalently attached to the 3' poly(A) tail of mRNA. Magnetic beads are used in many studies for this step [17, 21]. The next step is to reversely transcribe the RNA to cDNA and further fragment

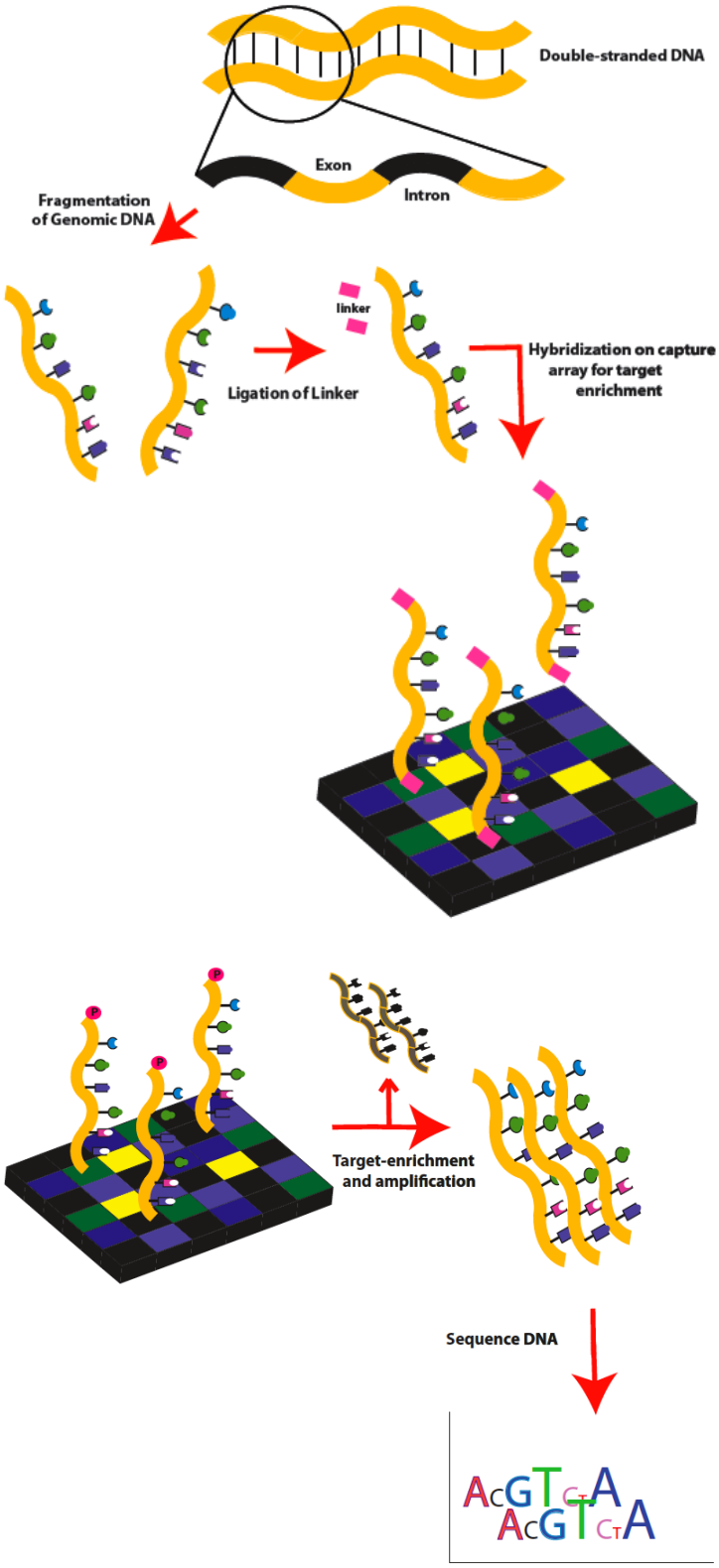


Figure 1.3 Exome sequencing workflow .

the cDNA to reach the desired length. The templates are then ready to be prepared for the sequencing. The workflow of RNA-Seq is shown in **Figure 1.4**.

4.4. ChIP-Seq

ChIP-Seq (chromatin immunoprecipitation sequencing) is used to identify the binding sites of DNA-associated proteins. ChIP-Seq is primarily used to study how DNA-associated proteins, such as transcription factors and histone, interact with DNA to regulate the gene expression, which plays an essential role in deciphering biological processes.

Some DNA sites interact directly with transcription factors or other proteins to form DNA-protein complexes, which can be isolated by chromatin immunoprecipitation with antibody against the protein of interest. Then the small pieces of DNA bound to the interested protein are ligated to oligonucleotide adaptors for the following sequencing (**Figure 1.5**).

5. Bioinformatics for NGS

Although the NGS technologies are developing at a rapid pace, the short read and the sheer scale data remains a significant challenge in data analysis.

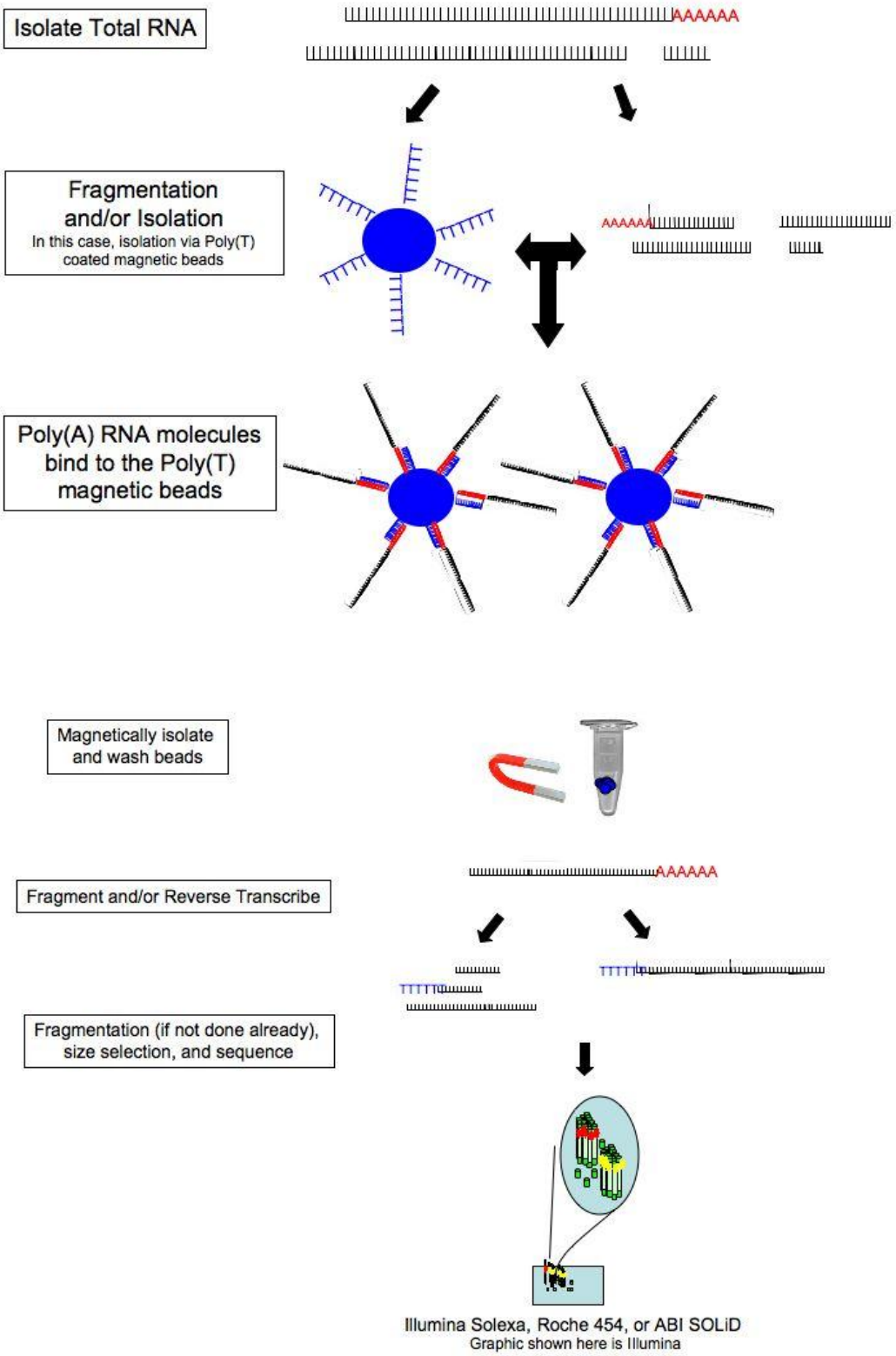


Figure 1.4 Workflow of RNA-Seq .

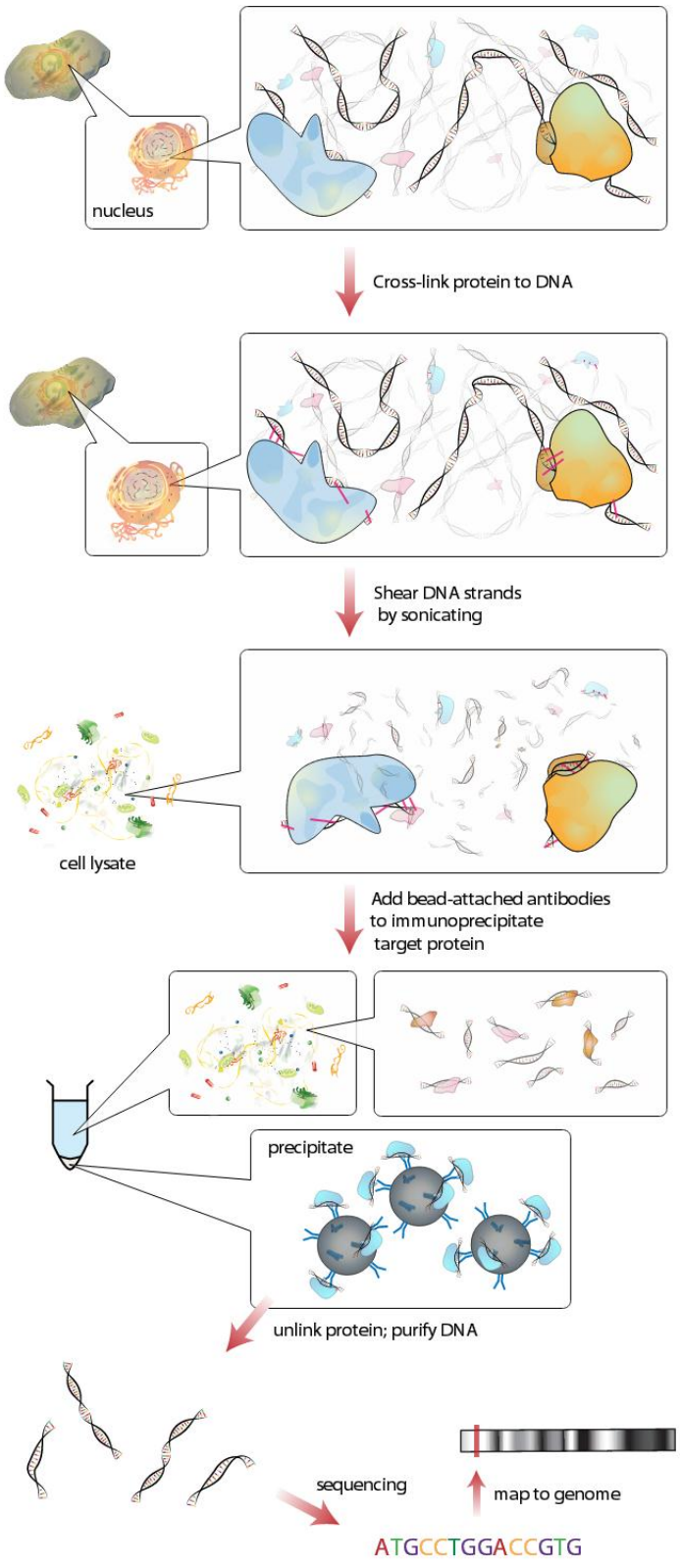


Figure 1.5 Workflow of RNA-Seq .

5.1. Sequencing error and quality score

The primary data from NGS consists of raw sequencing reads and the quality score for each base. The quality scores from different sequencing platforms cannot be compared directly, but all of them are Phred-like scores that are related to the sequencing error probabilities logarithmically. Different sequencing platforms generate various types of error. For the 454 platform, the length of each homopolymer is inferred from the observed fluorescence intensity, while the variance of the intensity for a specific homopolymer length is large. So 454 system has high error rate in insertion and deletion (indel) calls. For the Illumina platform, indels are rare. The major sequencing errors come from miscall, with a typical rate of ~1%. The SOLiD platform uses dinucleotide encoding scheme, in which each base is called twice. Thus the sequencing error rate in SOLiD is relatively smaller. For both Illumina and SOLiD platforms, base calling becomes less accurate towards the ends of reads. Depending on the platform, the error rate of NGS data ranges from several per cent to tenths of per cent. Reducing the sequencing errors is important to the assembly, alignment, variants detection and other downstream genomic analysis.

5.2. NGS analysis pipeline

Usually after the NGS reads are generated, the first step is either aligning the reads to a reference genome or doing *de novo* assemble, which is the basis of the analysis thereafter.

5.2.1. Alignment

Aligning the massive, short sequencing reads to the huge reference genome is a computationally complicated problem. Many alignment algorithms were developed and some the most popular ones are MAQ, BWA and bowtie. Illumina has developed their own aligner Eland, and SOLiD also developed Bioscope for their customers. There are some limitations to the alignment approaches. Errors often occur because of the ambiguous bases and the sequencing errors in the short reads. Multi-mapping are frequently observed when the reads are placed in the repetitive regions in the reference genome [24]. Moreover, the presence of gaps and misassemblies in the imperfect reference genome also leads to misalignment [25, 26]. Pair-end reads can resolve the misalignment for some repetitive regions if one read in the pair is unique to the genome.

5.2.2. *De novo* assembly

de novo assembly refers to aligning and merging the sequencing reads to reconstruct the original sequenced genome. It is important as the reference genome is lack for most species. Compared with alignment, assembly is computationally orders of magnitude slower and more memory intensive. *De novo* assembly has been successfully applied to assemble the bacterial genomes and mammalian bacterial artificial chromosomes [27-31]. However, the application to human genome remains a substantial challenge. The short read length usually leads to a lot of gaps, regions without reads aligned, and the sequencing errors often cause branching, resulting in poor assembly quality. Pair-end reads could partially resolve this problem and produce longer contigs by filling gaps in the consensus sequence.

5.2.3. Variants detection

Variants detection is one of the most important applications of NGS, with the challenge of separating the real variants from sequencing errors. Most variants detection methods use Bayesian algorithms to estimate the probability of calling a variant at a specific position. Variants detection has become more sophisticated and the further steps, such as local realignment around indels, quality score recalibration, and removal of duplicates, are usually implemented to improve the accuracy of variant calling. Once the variants are detected, they are typically annotated to predict the functional significance.

6. Significance and specific aims

Since the first introduction in 2005, NGS technologies have generated an incredible impact on genomic research. They have been broadly applied to many fields including genomic variation detection, gene expression and profiling, protein-DNA interaction, detection of aberrant transcription, small ncRNA discovery and profiling, genome annotation, and epigenomics. However, the development of methods for NGS data analysis is far away behind the advances of NGS technology itself. More efficient data analysis methods are required to establish pipeline for many applications before the analysis becomes routine. One major obstacle of data analysis is sequencing error, which affects the efficiency and accuracy of the analyzed results. Therefore, the purpose

of my study is to design data analysis methods that can minimize the effect from sequencing errors. The two aims are described as below:

1) **To design a robust and efficient error-correction method to correct the sequencing reads before the downstream analysis.** Most of the existing error-correction methods require high and uniform coverage, which does not fit some sequencing applications, such as single-cell sequencing, and mRNA-Seq. Here, we will design an error-correction method without any requirement for the sequencing coverage.
2) **To design a SNP calling method.** The existence of sequencing error usually results in low SNP-calling accuracy for low coverage data. Here we will design a new model to accurately detect the SNPs without requirement for the sequencing depth of the data.

CHAPTER TWO

MTM: An Error-Correction Method for High-Throughput Sequencing Data

1. SUMMARY

Background

NGS technologies produce massive amounts of short reads. The high sequencing error rate has become one of the major obstacles in sequencing data applications, such as *de novo* assembly and re-sequencing. Thus, error correction prior to data analysis is a critical step for the success of downstream analysis. Several error-correction methods have been developed in recent years. Most of those methods work with the assumption that sequencing reads are uniformly distributed. Although a few methods work well on the non-uniform data, they are computationally expensive and do not perform as well as the uniformity-specific methods on the uniform data. Therefore, a robust and efficient error-correction method is in urgent need.

Results

We report MTM, a new method for correcting errors in the NGS reads without the uniformity assumption. By using a mutating-testing algorithm, MTM outperforms the pervious error-correction methods with robustness, as well as higher positive predictive values and sensitivities. We also demonstrated the improvements of error correction with MTM on the mapping and variant detection.

Conclusions

MTM is a robust and efficient error correction method for NGS data, which can improve the quality of the results in the subsequent analysis. It is implemented in C++ and has been released as a software package downloadable at https://sites.google.com/site/mtmerrorcorrection/

2. INTRODUCTION

NGS technology developed in the last decade provides monumental increase in speed and volume, taking biological and biomedical research to a whole new level. For example, tumour samples from thousands of patients (TCGA) have been sequenced in order to discover the complete set of oncogenes and tumour suppressor genes. The 1000 Genome Project is using sequencing to establish by far the most detailed catalogue of human genetic variations [32]. Thus far, the applications are mostly confined to the organisms whose genomes have been sequenced. A more exciting possibility is to expand the new sequencing capacity to the study of genome biology of previously unexplored organisms. Meanwhile, the Genome 10K Project plans to sequence and assemble the genomes of 10,000 vertebrate species [33].

Compared to the traditional shotgun methods [34], the NGS techniques generate a much larger set of shorter reads with higher error rates, which challenge the downstream analysis tools. Sanger reads, typically 700-1000 bp long, were assembled using an overlap-layout-consensus approach, which would be too slow when applied to the NGS data. Thus, a new method has been developed using *de Bruijn* graph [35] that is much faster to compute large amount of sequencing data [36]. However, methods based on *de Bruijn* graph are highly sensitive to sequencing errors [37], which cause branching in the graph and greatly increase the computational cost. Therefore correcting sequencing errors before assembly is a build-in feature of almost all *de Bruijn* assemblers. Re-sequencing is another important application of NGS technology. Reads are aligned to the reference genome by allowing up to a fixed number of mismatches caused by either polymorphisms or sequencing errors [38]. Therefore, some reads are

difficult to be mapped to the reference genome due to the sequencing errors, especially in the polymorphism-rich regions. Pre-processing the reads to eliminate sequencing errors will improve the mapping ability. Subsequently, the sensitivity and specificity of variant detection will be improved as well.

The general idea behind error correction methods is to align all the reads that cover the same genome locations, and identify the erroneous base using the high coverage of the NGS technology. As the reference genome is unknown, the reads from the same genome location refer to the reads sharing the same subsequence of a fixed length *k*, called *k*-mers [39]. Established error correction methods can be classified into three types - *k*-spectrum based [28, 36, 40-47], suffix tree/array-based [48-50] and multiple sequence alignment (MSA) -based methods [51, 52]. A common approach of error correction is to use the frequency of *k*-mers to separate them into trusted *k*-mers and untrusted *k*-mers. *k*-mers with low frequency usually represent sequencing errors, while *k*-mers with high frequency are likely to occur in the genome. When the sequencing coverage is high and uniform, the distributions of trusted *k*-mers and untrusted *k*-mers are separated. By choosing a right threshold, they can be separated very well [28, 36, 43, 46, 47]. However, these methods do not work well when the data does not cover the genome uniformly. For example, data from transcriptome sequencing (RNA-Seq) and single-cell sequencing, the coverage of which are dramatically uneven. Also, when the DNA is from environmental samples and cannot be cloned, the amount of starting materials is small. Using above methods to pre-process these data is not effective and results in loss of real reads. Another popular approach is based on Hamming graph. *k*-mers within a small Hamming distance are grouped together, and the

k-mers with low frequency are corrected to the high-frequency *k*-mer [44, 48-50, 53, 54]. Although some of these methods work well for non-uniform data, but they can be computationally expensive and do not perform as well as the methods based on *k*-mers frequency on uniform data.

In this paper, we present a new method MTM to correct sequencing errors, without the assumption on the uniformity of the data. MTM is similar in spirit to the Hammer graph methods. It assumes that the *k*-mer population is consisted of high frequency error-free *k*-mers and low frequency erroneous *k*-mers, and that an erroneous *k*-mers can be linked to an error-free *k*-mer by a small number of point mutations. It further assumes that the ratio of the frequencies of the linked *k*-mers is consistent to the sequencing error rate. MTM outperforms the previous non-assumption methods on both uniform and non-uniform data, and achieved similar performance to uniformity-specific methods on uniform data. MTM is efficient on dealing with large datasets or data with high error rates without the limitation on memory. Moreover, MTM allows users to choose a large range of *k*-mer length, providing more flexibility for the downstream analysis. Finally, we explored the impact of error correction with MTM on mapping ability and variant detection. After error correction, we were able to map more reads to the reference genome, and identify more variants while remaining the same precision.

3. MATIRIALS & METHODS

3.1. Data

To test the effectiveness of MTM on both uniform and non-uniform datasets, we used two data sets: 1) Single-cell data with highly non-uniform coverage [55]. The reads were amplified from a single-cell of *E.coli* K12 MG1655 and sequenced by the Illumina GAII pipeline (lane 1). The 100 bp long reads with ~600× sequencing depth result in 94 blackout regions and totally 116 kbp with 0 or 1 coverage. 2) Normal multi-cell data with uniform coverage (ERX002508). The data were also generated from *E.coli* K-12 MG1655 and sequenced by the Illumina GAII pipeline with the same coverage and read length.

Since the sequencing accuracy is low at the beginning and end of a read, prior to working with the datasets, we trimmed the reads by only keeping the longest region of a read that are longer than *k* (size of oligonucleotide) and do not contain ambiguous bases or bases with quality score lower than QUAL_LOW. In this study, we let QUAL_LOW equal 3.

3.2. Statistical model

Two *k*-mers S_1 and S_2 with multiplicities m_1 and m_2 ($m_1 > m_2$) have one nucleotide difference in the sequences. Then S_2 may either be sequenced from S_1 with one sequencing error or be a true repeat of S_1 .To distinguish the sequencing error from repeat, we tested if the percentage of S_2 is consistent with the average sequencing error rate. Define q as the average sequencing error rate and p as the percentage of *k*-mer S_2. Then

$$H_0: p = q$$

$$H_1: p \neq q$$

Since $m_2 \sim Binomial\ (m_1 + m_2\ ,\ p)$, the likelihood function is

$$L(p|m_1, m_2) = \binom{m_1 + m_2}{m_2} p^{m_2}(1-p)^{m_1}$$

When $p = \frac{m_2}{m_1+m_2}$, $L(p|m_1, m_2)$ is maximized. Therefore, the likelihood ratio test statistics is

$$D = -2\log \frac{\sup_{p=q} L(p|m_1, m_2)}{\sup\ \ L(p|m_1, m_2)} = -2\log \frac{\binom{m_1 + m_2}{m_2} q^{m_2}(1-q)^{m_1}}{\binom{m_1 + m_2}{m_2} \left(\frac{m_2}{m_1 + m_2}\right)^{m_2} \left(\frac{m_1}{m_1 + m_2}\right)^{m_1}}$$

The distribution of *D* is close to χ^2 distribution. By defining a significance level α, null hypothesis is rejected if $D > c_\alpha$, where $\mathrm{P}(D > c_\alpha) = \alpha$. The rejection of null hypothesis indicates that *k*-mer S_2 is not the erroneous form of S_1.

3.3. Algorithm

MTM detects and corrects the sequencing errors by converting the reads to *k*-mers and distinguishing trusted *k*-mers from erroneous *k*-mers, then using the trusted *k*-mers to correct the original reads. Because sequencing error is small, the error free sequences should occur more frequently than the sequences with errors. Thus, *k*-mers with higher multiplicities are more trustable. MTM mutates each *k*-mer and searches for its close *k*-mers with lower multiplicities. If a match is found, the above statistical model is applied to determine if the matched *k*-mer is likely to be an erroneous form of the original *k*-mer. MTM consists of the following steps:

1) **Counting *k*-mers**: the first step of MTM is cutting the trimmed reads to *k*-mers and counting the occurrences of all *k*-mers. A read with length *L* produces $L - k + 1$ *k*-mers.
2) **Sorting *k*-mers**: *k*-mers are sorted by their multiplicities from high to low.
3) **Mutating and testing**: Starting from the *k*-mer with highest multiplicity, we work on all the *k*-mers in the order sorted in step 2. For each *k*-mer S_o, we mutate one nucleotide at a time to generate a new *k*-mer S_m with only one substitutional difference to the original *k*-mer S_o. A *k*-mer can be mutated to $3k$ new *k*-mers as shown in **Figure 2.1(a)**. Then we search each new *k*-mer S_m in the *k*-mer list with multiplicities lower than S_o. Once a match is found, the above statistical model is applied to test if S_m is the result of erroneous sequencing of S_m. If the testing result shows that S_m is erroneous, S_m will be corrected to S_o at the end of this step. Meanwhile, a table is generated to record erroneous *k*-mers and their corresponding corrected *k*-mers. Since the mutation and correction processes are accumulative, correction could happen between two *k*-mers with more than one bases difference, such as the k-mers S_a and S_c in **Figure 2.1(b)**. Moreover, although error free *k*-mers occur more frequently than erroneous *k*-mers, high multiplicity does not guarantee that a *k*-mer is trustable. It is because the sequencing coverage are often uneven, which is especially true when amplifying from a small amount of starting materials and in RNA-Seq. Even when de novo sequencing is performed at optimal condition, the coverage often distributes in a wide range. A non-negligible portion of the genome have high coverage. Therefore, a high multiplicity k-mer can be the result of sequencing

error from a k-mer with even higher multiplicity, just like S_b in **Figure 2.1(b)**, which is erroneous form of S_a.

4) **Correcting sequencing reads**: Using the correct *k*-mers list and the correction table mentioned in step 3, we map the correct *k*-mers back to the original reads. If no correct *k*-mer is mapped to a position, the original base in the read is kept. Otherwise, the consensus nucleotide is the base in the final correct sequence of the read (**Figure 2.2**).

3.4. Parameters

MTM has three parameters: 1) α, significance level of likelihood ratio test for determining if the mutated *k*-mer is trustable. Setting α too large may reduce the power of correction, remaining a lot of erroneous *k*-mers after correction. Whereas setting α too small may reduce the sensitivity, thus losing some true *k*-mers. We optimized α as 0.1% in our method (**Figure 2.3**). 2) *k*, the length of oligonucleotide MTM works on. Like the trades-offs with *k*-mer size in genome assembly, too small of a *k* results in a high probability that one *k*-mer in the genome would be similar to another *k*-mer in the genome with only one nucleotide substitution, making the situation more complicated. Too large of a *k* results in low *k*-mer coverage and reduces the accuracy of algorithm. We designed two versions of codes for MTM – binary version and string version. Binary version is faster than the string version, while it has a limitation for *k* ($k \leq 32$). 3) *mulcutoff*, the threshold of multiplicity below which the *k*-mers will be removed. Varying the value of *mulcutoff* results in a trade-off between sensitivity and specificity.

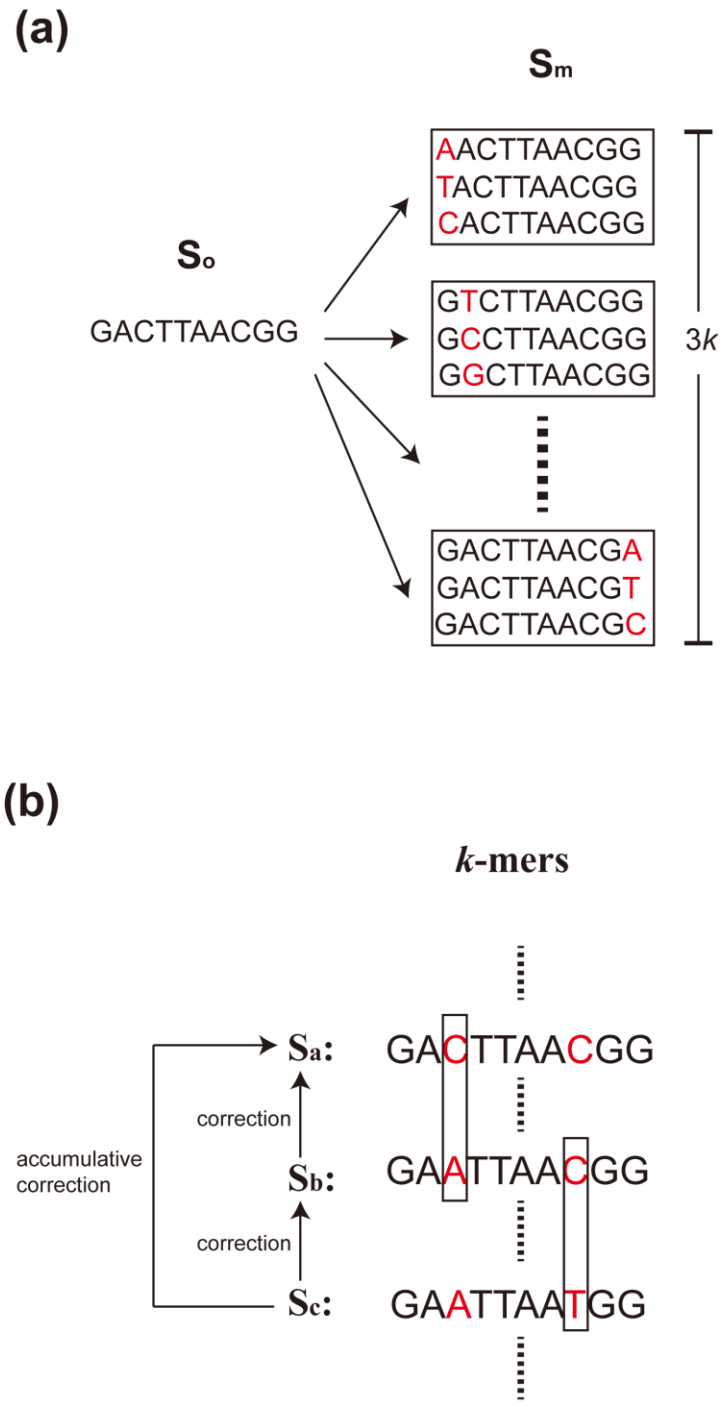


Figure 2.1. Illustration of *k*-mer mutation and accumulative correction. (a) One *k*-mer is mutated to 3*k* new *k*-mers with one nucleotide substitution. (b) This is an example of accumulative correction. We mutate *k*-mer S_a by one nucleotide and find a match S_b. The statistical test shows S_b is the result of sequencing error from S_a. So S_b is corrected to S_a. Similarly, then we do mutation to S_b and find out that S_c can be corrected to S_b. Therefore, S_a is the original correct form of S_c.

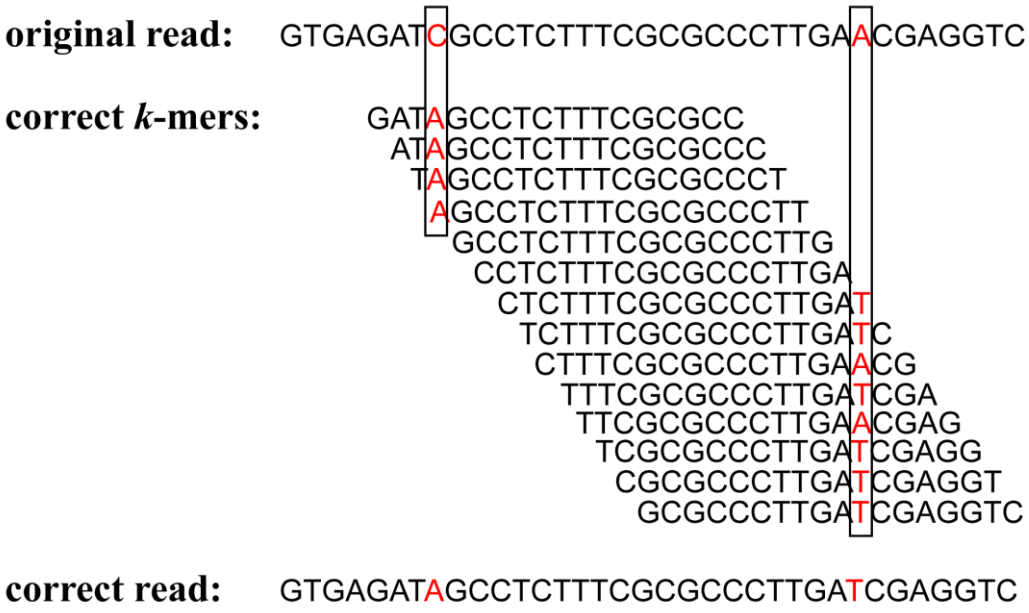


Figure 2.2. Correction of original sequencing reads with trusted *k*-mers. This is an example indicating the way that the sequencing reads are corrected based on the correct *k*-mers. We mapped *k*-mers back to the original read, and correct the sequence of the original read to the consensus of *k*-mers. The region with no *k*-mers mapped is saved.

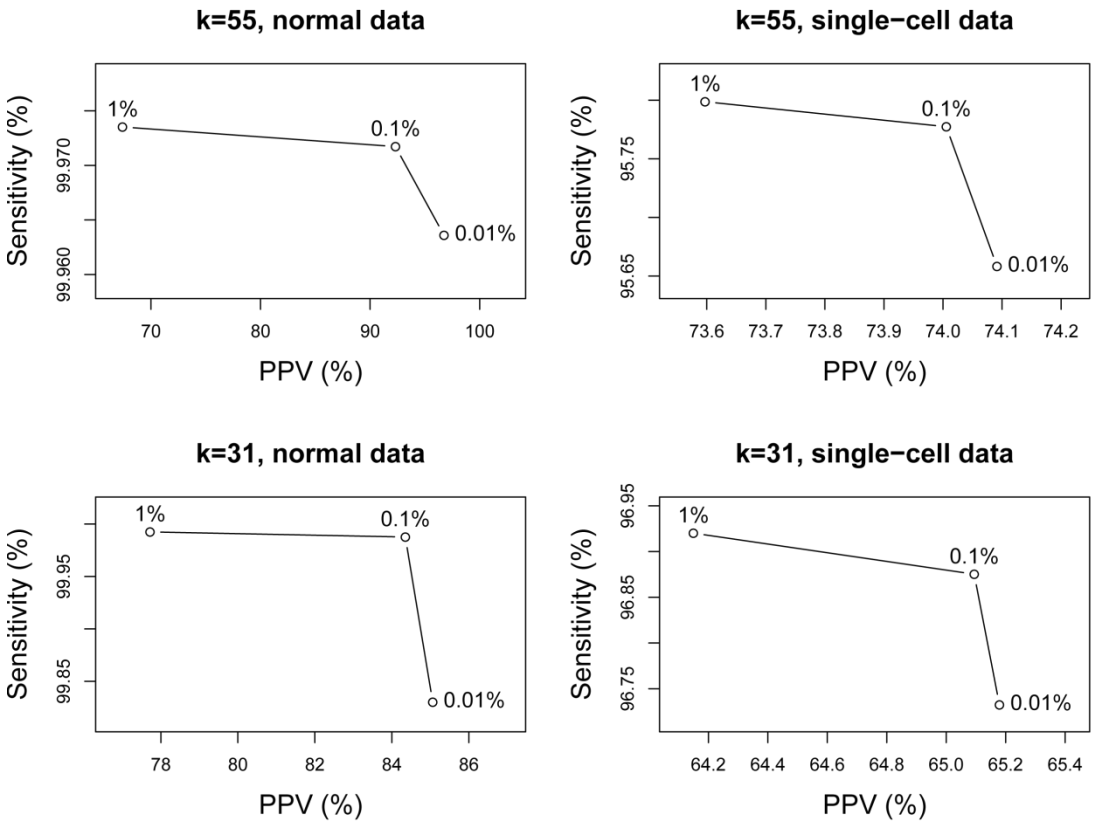


Figure 2.3. Sensitivity ~ PPV plots of MTM with different α values. We measured the sensitivities and PPVs by choosing different α values. α value is labeled next to the curves. When α changes from 0.1% to 0.01%, the sensitivities decrease. So on balance, 0.1% is the recommended value for α.

4. RESULTS

MTM corrects errors in the sequencing reads. To achieve this goal, MTM converts the reads to *k*-mers and separates the error-free *k*-mers from the erroneous *k*-mers, and then replaces the *k*-mers from the original reads with error-free *k*-mers. Therefore, finding out the error-free *k*-mers is the critical step of MTM. To assess MTM's ability, following Medvedev *et al.*[45], we measured the data's sensitivity and positive predictive value (PPV) with respect to the reference *E.coli* genome. Sensitivity is measured by the percentage of *E.coli* genome's *k*-mers that are present in the dataset. PPV is the percentage of data's *k*-mers that are present in the *E.coli* genome.

To test the performance of MTM on different *k* values, two different values of *k* (*k* = 31 and *k* = 55) were used to run MTM for each dataset. We ran the binary version program for *k* = 31 and ran the string version program for *k* = 55. To increase the quality of the reads, we trimmed the data before running MTM, and about 85% - 91% of the data were preserved. The trimming step did not affect the high sensitivities of the data, which are 99.98% (*k* = 55) and 99.99% (*k* = 31) for the normal data, and 97.04% (*k* = 55) and 97.72% (*k* = 31) for the single-cell data. Error correction by MTM dramatically reduced the percentage of erroneous *k*-mers, especially when the *mulcutoff* is larger than one (**Figure 2.4**).

4.1. Comparison to other methods

Quake is a program that is superior in detecting as well as correcting sequencing errors in the data with high and uniform coverage [43]. Quake works by weighting the *k*-mer multiplicities with quality values and then modelling the histogram of weighted

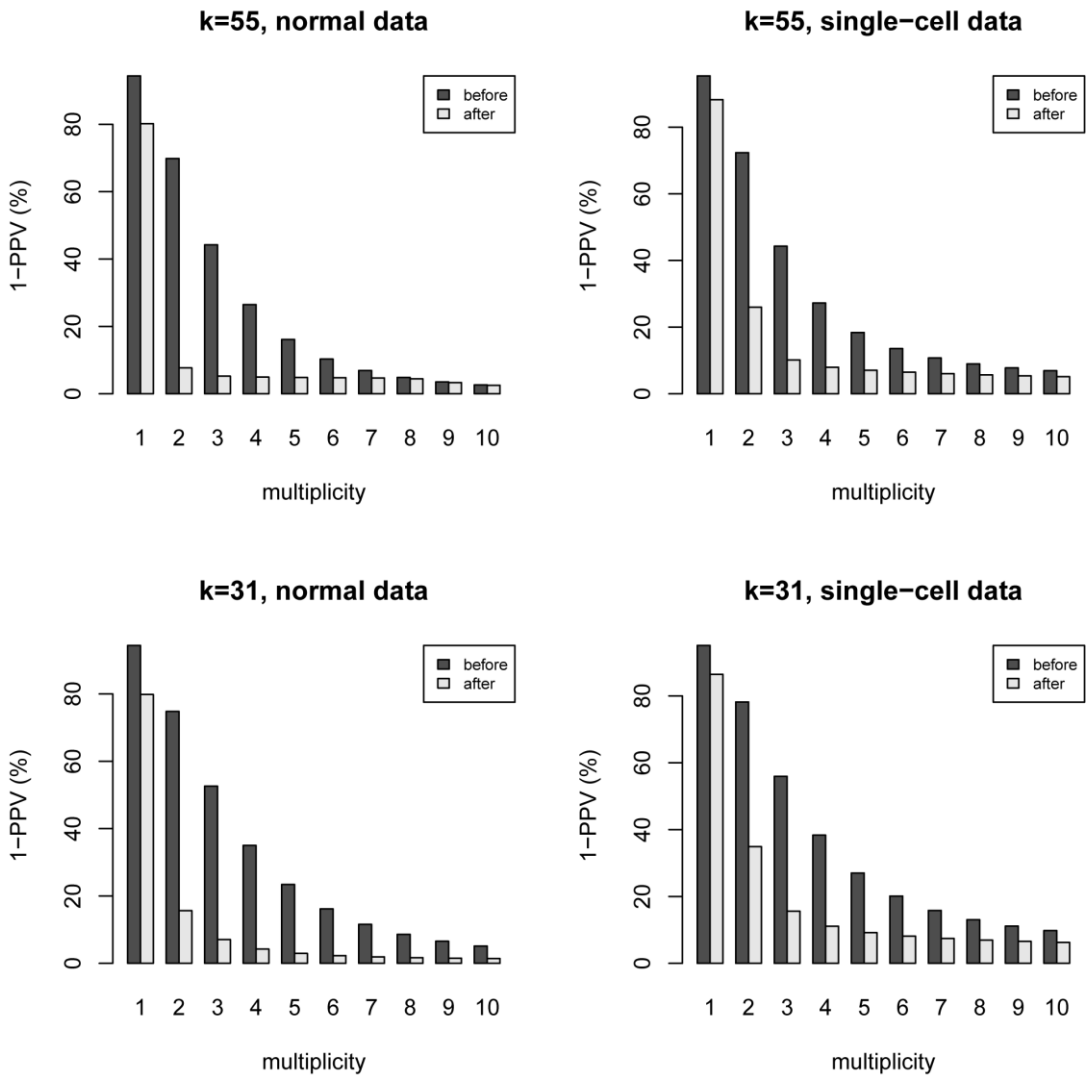
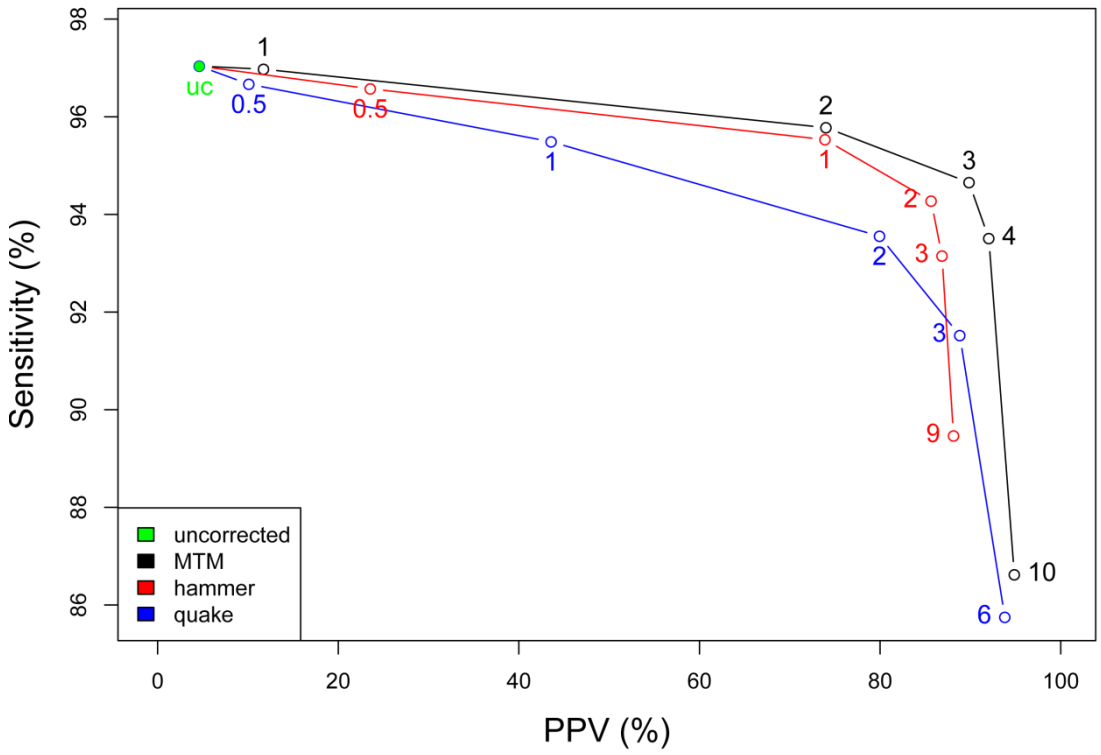


Figure 2.4. Error correction by MTM removed the majority of the erroneous *k*-mers. We measured the PPV of normal and single-cell data by varying *mulcutoff* in the MTM. After the error correction by MTM, the percentage of erroneous *k*-mers dramatically decreased, especially when $k \geq 2$. The vertical axis is 1-PPV, which is the percentage of data's k-mers that are not present in the *E.coli* genome. Black color represents the results before the correction, and the grey color represents the results after correction.

multiplicities as at mixture of two distributions to choose an appropriate cutoff between error-free *k*-mers and erroneous *k*-mers. Quake is not able to find the cutoff for the single-cell data, so we manually tried different values as the cutoff and compared the results with MTM (**Figure 2.5**). MTM works better than Quake because its curve is closer to the top right corner, which indicates higher sensitivity and PPV. Based on the plots, we choose 4 as the cutoff for MTM, because by using the cutoff of 4 MTM is able to obtain high PPV with a negligible decrease of sensitivity. Accordingly, the cutoff of Quake is set to 3. The number of the comparison results is showed in **Table 2.1**. MTM outperformed Quake for single-cell data with both higher PPV and sensitivity, and achieve comparable performance for normal data. When the cutoff is set to 4 for the normal data, MTM get higher sensitivity and lower PPV than Quake. When the cutoff is increased to 9, MTM win out with higher PPV and the same sensitivity.

Another method called Hammer was recently developed, which is popular for its good performance on correcting non-uniformly distributed data [45]. Based on a combination of Hamming graph and a probabilistic model, Hammer identifies the clusters for similar *k*-mers and generates a consensus *k*-mer as the error-free k-mer for each cluster. We also compared MTM with Hammer, and the results showed that MTM improved Hammer on both single-cell data and normal data. **Figure 2.5** shows that MTM gets higher PPV and sensitivities at various cutoff points for single-cell data. The number of the comparison results is shown in **Table 2.1**, in which the *singletonCutoff* of Hammer is set to 3 according to the plots. With *k*=55, MTM retained ~16K more true *k*-mers, while reduced the erroneous *k*-mers by ~258K for single-cell data. Similar results are obtained with *k*=31.

(a) k = 55


Sensitivity (%)
PPV (%)
uncorrected
MTM
hammer
quake
uc
1
2
3
4
10
6
9
0.5
0.5
1
1
2
2
3
3

(b) *k* = 31

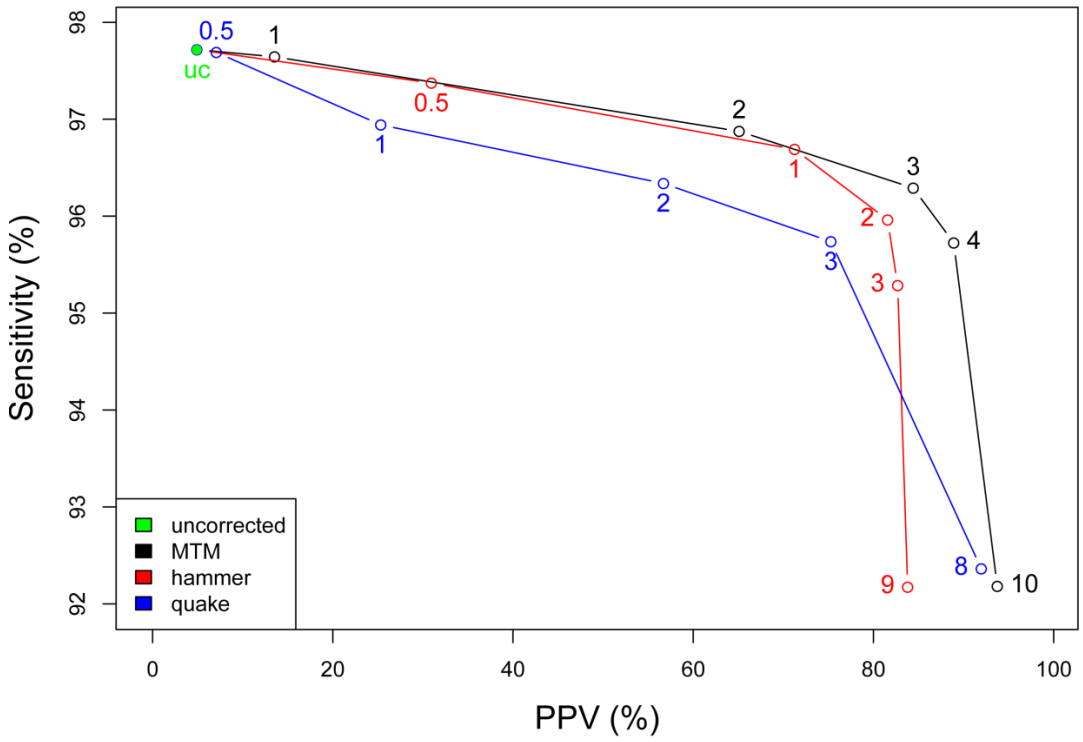


Figure 2.5 Sensitivity ~ PPV plots with different error correction tools.
For single-cell data, we plotted the sensitivity and PPV with different error correction tools by varying the cutoff values. MTM performs better than Quake and Hammer because its curve is closer to the upper right corner, where the sensitivity and PPV are higher. Cutoff values were labelled next to the curves with the same color as the curve. For MTM, the cutoff is *mulcutoff*, *k*-mers with the multiplicity below which are discarded. For Hammer and Quake, the cutoff is weighted multiplicity. "uc" represents "uncorrected", which is the value before correction.

Table 2.1 Comparison of MTM results with other tools. MTM outperforms Hammer for both single-cell data and normal data. MTM also shows better performance than Quake for single-cell data, and comparable results for normal data. The cutoff for MTM is the value of parameter *mulcutoff*, *k*-mers with the multiplicity below which were discarded. The cutoff for Hammer is the value of parameter *singletonCutoff* in Hammer, which is the threshold for the weighted multiplicity. For Quake, the cutoff is also based on the weighted multiplicity, and it is calculated automatically by the program for normal data.

(a) Comparison results with $k = 55$.

		distinct *k*-mers	**correct *k*-mers**	**PPV (%)**	**sensitivity(%)**
Single-cell	Before correction	9,602,1803	4,430,156	4.61	97.04
	MTM (cutoff = 4)	4,638,400	4,268,971	92.04	93.51
	Hammer (cutoff = 3)	4,896,130	4,252,619	86.86	93.15
	Quake (cutoff = 3)	4,703,965	4,178,234	88.82	91.52
normal	Before correction	81,042,139	4,564,457	5.63	99.98
	MTM (cutoff = 4)	4,801,122	4,563,700	95.05	99.96
	Hammer (cutoff = 3)	4,857,315	4,562,884	93.94	99.94
	Quake	4,725,473	4,562,648	96.55	99.94
	MTM (cutoff = 9)	4,719,529	4,562,509	96.67	99.94

(b) Comparison results with k = 31.

		distinct *k*-mers	**correct *k*-mers**	**PPV (%)**	**sensitivity(%)**
Single-cell	Before correction	90,598,084	4,450,268	4.91	97.72
	MTM (cutoff = 4)	4,903,206	4,359,514	88.91	95.72
	Hammer (cutoff = 3)	5,247,985	4,339,490	82.69	95.28
	Quake (cutoff = 3)	5,792,076	4,360,114	75.28	95.74
normal	Before correction	81,128,182	4,553,932	5.61	99.99
	MTM (cutoff = 4)	4,755,026	4,553,849	95.77	99.99
	Hammer (cutoff = 3)	4,984,462	4,550,950	91.30	99.93
	Quake	4,599,644	4,553,533	99.00	99.98
	MTM (cutoff = 19)	4,597,788	4,553,474	99.04	99.98

The superiority of MTM over Quake and Hammer becomes more significant when the data quality is low. In the above analysis, we trimmed the data by setting the QUAL_LOW to 3, which grabbed the longest region in a read that is longer than *k* and does not contain ambiguous base or base with quality score lower than QUAL_LOW. In this part, we set the QUAL_LOW to 2. In other words, we kept all the low quality bases in the dataset. As a result, about 8% more data was preserved on average. We plotted the sensitivities and PPVs from different methods (**Figure 2.6**). Compared to **Figure 2.5**, the distances between the MTM curve and other two curves are larger, indicating that the improvement of MTM on Quake and Hammer is more significant.

4.2. Parameters selection

Optimizing the parameters could lead to better performance. The first parameter is the length of *k*-mer. Longer *k*-mers tend to have more sequencing errors and consequently reduce the chance that they can be reached from the trusted *k*-mer within a finite number of mutations. As a result, MTM is expected to work better with smaller *k*. **Figure 2.7** shows the results of the single-cell data with $k = 55$ and $k = 31$ separately. The curve for $k = 31$ is better because it is closer to the top right corner. This figure indicated that the algorithm prefers smaller *k*. However, if the *k*-mer is too small, there is a greater risk for an unrelated *k*-mer to accidentally match to the genome. Thus, one should balance the pros and cons to choose an appropriate *k*-mer length.

α is the significance level of the statistical test that is used to distinguish the trusted *k*-mers from erroneous *k*-mers. We measured the sensitivity and PPV by varying

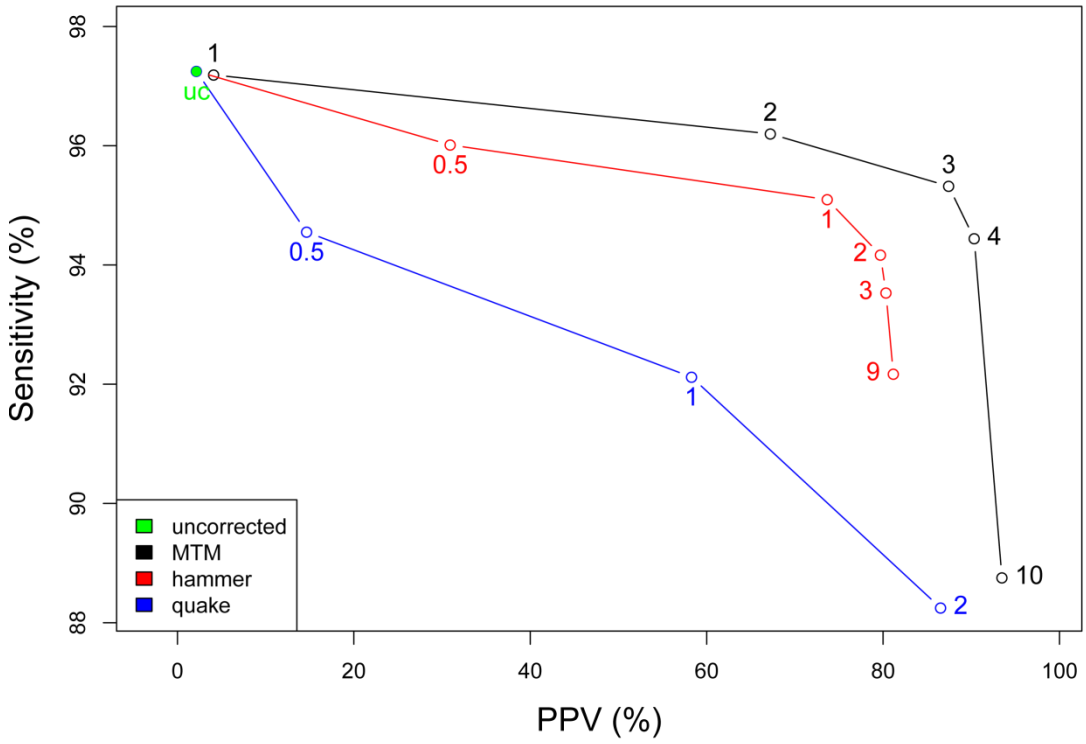


Figure 2.6 Sensitivity ~ PPV plots with different tools for low quality data (*k*=55). Single-cell data in this figure was trimmed with QUAL_LOW=2. Namely, the low quality bases were not trimmed off, and only ambiguous bases were removed. With the correction results, we plotted the PPV-sensitivity curves for MTM, Hammer and Quake. Compared to Figure 2, the distances between MTM curve and other two curves are much larger, indicating that the improvement of MTM on Hammer and Quake is more significant. Cutoff values were labelled next to the curves with the same color as the curve. For MTM, the cutoff is *mulcutoff*, *k*-mers with the multiplicity below which are discarded. For Hammer and Quake, the cutoff is weighted multiplicity. "uc" represents "uncorrected", which is the value before correction.

α value (**Figure 2.3**). When the value of α changes from 1% to 0.1%, the PPV increases, while the sensitivity only has a slight decrease. However, the sensitivity decreases significantly when α changed from 0.1% to 0.01%. So considering the balance of sensitivity and PPV, 0.1% is recommended and is used in our analysis.

4.3. Improvement on mapping ability

Aligning the sequencing reads to the reference genome is the first step in the application of short reads. Mapping the short reads is usually achieved by allowing a fixed number of mismatches [38]. In other words, if the number of sequencing errors in a read is larger than the fixed number, the read will not be mapped. Thus, by performing error correction before mapping, the mapping ability will be improved. We mapped the short reads to their sequencing template *E.coli* K12 before and after the error correction with MTM. We used two different modes of Bowtie to do the mapping [56]. Firstly, We mapped the reads with Bowtie's default mode (-n mode), which allows no more than 2 mismatches in the first 28 bases, and the sum of the Phred quality values at all mismatched positions no more than 70. Then, we also did the mapping by allowing up to 2 mismatches for the whole read using the -v mode of Bowtie. The results are showed in **Table 2.2(a)**. In -n mode, about 50-70K additional reads were mapped after the error correction with MTM. And about 220-490K additional reads were mapped after the error correction with -v mode. The results clearly indicate the benefits of error correction on mapping, especially for the -v mode, which allows a few of mismatches in the whole read.

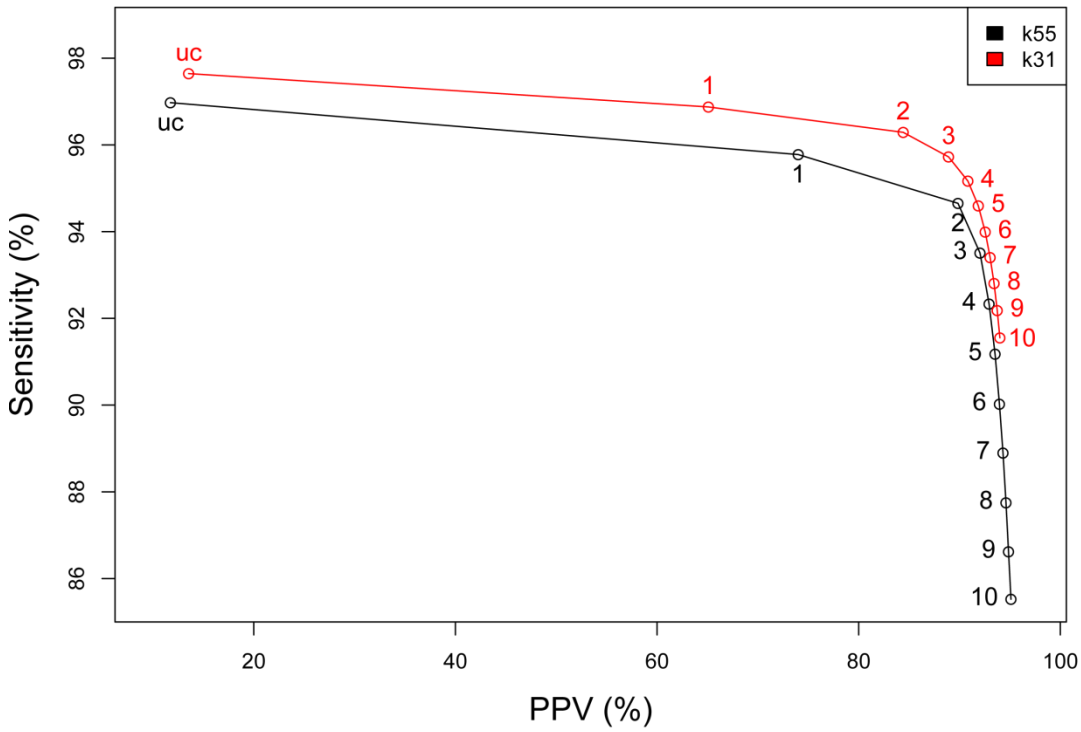


Figure 2.7 Sensitivity ~ PPV plots of MTM with different *k* values. We showed the sensitivity ~ PPV plots for k=55 and k=31 by varying the value of *mulcutoff*. Apparently, with k=31, MTM works slightly better. The *mulcutoff* values were labelled next to the curves with the same color as the curve. "uc" represents "uncorrected", which is the values before correction.

Table 2.2 Mapping results comparison. We used Bowtie to map the reads in two different modes: -n mode and -v mode. In -n mode, up to 2 mismatches were allowed in the first 28 bases, and the sum of Phred quality of all mismatches should not exceed 70. In -v mode, only up to 2 mismatches were allowed in the whole reads.

(a) *E.coli* K12 as the reference genome.

	n-mode		**v-mode**	
	before	**after**	**before**	**after**
lane1 (k=55)	24.92M (93.33%)	24.97M (93.51%)	24.56M (91.98%)	24.78M (92.81%)
normal (k=55)	26.72M (99.42%)	26.77M (99.61%)	26.17M (97.37%)	26.55M (98.79%)
lane1 (k=31)	26.64M (93.52%)	26.71M (93.76%)	26.26M (92.16%)	26.59M (93.33%)
normal (k=31)	27.82M (99.41%)	27.89M (99.69%)	27.25M (97.37%)	27.74M (99.13%)

(b) *E.coli* 536 as the reference genome

	n-mode		**v-mode**	
	before	**after**	**before**	**after**
lane1 (k=55)	13.43M (50.31%)	13.64M (51.09%)	12.27M (45.96%)	12.96M (48.52%)
normal (k=55)	13.63M (50.73%)	14.03M (52.21%)	12.70M (47.27%)	13.65M (50.79%)
lane1 (k=31)	14.66M (51.47%)	14.91M (52.35%)	13.43M (47.13%)	14.20M (49.83%)
normal (k=31)	14.41M (51.48%)	14.87M (53.14%)	13.43M (48.01%)	14.52M (51.88%)

In most cases, the genome being sequenced is different from the reference genome due to polymorphisms. The coexistence of SNPs and sequencing errors results in difficulties in mapping the reads to the reference genome, especially to the SNP-rich regions. So eliminating sequencing errors should benefit more when the reference genome is different from the sequenced genome. We used a related genome *E.coli* 536 [GenBank: NC_008253] as the reference genome to do the same mapping (**Table 2.2 (b)**). About 210-460K additional reads were mapped in -n mode, and about 690-1090K additional reads were mapped in -v mode, which demonstrated that error correction by MTM improves the mapping ability more significantly when the reference genome is different from the sequenced genome.

4.4. Improvement on SNPs calling

Variants detection is an important application of NGS. By correcting the sequencing errors before identifying the variants, mismatches between the aligned reads and reference genome can be reduced. Therefore the SNPs-clustered regions will be able to be mapped, resulting in more SNPs to be identified. To explore the benefit of error correction with MTM on variants detection, we used the same method that Quake used [43]. To call SNPs, we used the data that are sequenced from *E.coli* K12 but aligned to a relative genome *E.coli* 536 to detect SNPs with SAMtools [57]. To calculate the recall and precision statistics of the identified SNPs, we aligned the *E.coli* K12 genome and *E.coli* 536 genome with the *dnadiff* utility in MUMmer [58], and used the numerated SNPs as the gold standard. The results are showed in **Table 2.3**. After error correction with MTM, we discovered more SNPs. In both -n mode and -v

Table 2.3 - SNP calling. We called SNPs using the mapping results from Table 2.2 (b) with SAMtools. The SNPs were validated by comparing them with the SNPs from the alignment of *E.coli* K12 genome and *E.coli* 536 genome. Recall is the fraction of identified SNPs in the true SNPs. Precision is the fraction of true SNPs in the identified SNPs.

		SNPs		**Recall**		**Precision**	
		before	**after**	**before**	**after**	**before**	**after**
k=55	lane1, -n mode	68,270	68,932	0.621	0.627	0.991	0.991
	lane1, -v mode	56,287	56,916	0.512	0.518	0.992	0.992
	normal, -n mode	67,472	68,833	0.615	0.627	0.993	0.993
	normal, -v mode	61,044	61,830	0.556	0.564	0.994	0.993
k=31	lane1, -n mode	72,622	73,372	0.659	0.666	0.989	0.989
	lane1, -v mode	62,166	63,041	0.565	0.573	0.991	0.991
	normal, -n mode	71,809	73,277	0.653	0.667	0.992	0.992
	normal, -v mode	65,982	67,007	0.601	0.610	0.993	0.993

mode, the recall increased, while the precision did not change, indicating that the newly discovered SNPs are reliable.

5. DISCUSSION

The high throughput and low cost of NGS technologies produce a revolution in genome research. However, sequencing errors mislead and complicate the analysis of sequencing reads. Thus, preprocessing the sequencing reads to eliminate the sequencing errors is critical for improving the quality of downstream analysis. Despite the success of many error-correction tools, there is a lack of an efficient tool without limitations on the reads coverage, *k*-mer length and memory size. Most error-correction tools rely on the uniformity of the reads coverage [39], which is impossible for transcriptome sequencing and single-cell sequencing. Hammer is an alternate method working without the uniformity assumptions. MTM, the tool we present here, is also an assumption-free method. It outperforms the previous methods with higher PPV and sensitivity. Especially, when the quality of the data is low, MTM shows more significant superiority.

There is no limitation on the *k*-mer length, which is a strength of MTM. We found that the processing time and required memory of Hammer increases significantly when *k* is small, making it not applicable in genome assembly. Also, the correction and reconstruction steps of Quake do not support large *k* due to memory limitations (19-mers require 32 GB). In contrast, the flexibility of *k* selection makes MTM easy to satisfy various needs. For example, long *k*-mers have the potential to be used to

distinguish the repetitive regions in eukaryotic genomes. We also provide a binary option for MTM, which greatly speeds up the analysis when *k* is no more than 32.

Rather than correcting *k*-mers, MTM detects putative erroneous *k*-mers and removes them. In other words, using MTM, no new *k*-mers is introduced and all the *k*-mers used to reconstruct the reads come directly from the original reads. As a consequence, the error correction by MTM is not able to increase the sensitivity of the data. However, without creating new *k*-mers, MTM eliminates the risk of introducing false positive. Error-correction methods are designed to mainly target hyplotype genome sequencing so far, including MTM. So extending the current MTM implementation on processing highly repetitive eukaryotic genomes is part of our future work. This can be achieved by extending the *k*-mer on both ends and searching their consensus *k*-mers. The surrounding sequencing should have consistent consensus *k*-mers with the original *k*-mer. Furthermore, like most error-correction algorithms, MTM only targets on substitution errors, which is the main source of errors in Illumina sequencing platform. The emergence of new platforms, such as PacBio sequencer and Ion Torrent that are abundant of indels, challenge MTM. So adopting insertions and deletions in the error correction process is our next target.

CHAPTER THREE

A New Method of Discovering Genome-Wide SNPs from Next-Generation Sequencing Data

1. SUMMARY

Background

The advance of NGS technology provides an unprecedentedly efficient way to comprehensively catalogue the human genetic variants. Ideally, SNPs can be effectively detected by counting the allele frequency. However, the high sequencing error rates complicate the situation, especially for low coverage (< 5×) data. Recently, many NGS studies are based on data with low to medium coverage (< 20×). The increasing demand for sequencing more samples suggests that the low or medium coverage sequencing may be the most common and cost-effective design. Therefore, accurate SNP calling methods without the limitation in the sequencing coverage is important to future genetic studies.

Results

To improve the existing methods, we developed a Bayesian-based approach for calling SNPs without requirement for the sequencing depth. We successfully applied this approach to identify the SNPs in prostate cancer cell line PC-3 and colon cancer cell lines RKO and SW48, whose mutation status are unknown. Our method outperforms the existing methods by identify more SNPs while maintaining higher dbSNP rates, especially for the low coverage PC-3 data. Eventually, we identified 107 potential causal genes for PC-3. For RKO and SW48 cell lines, 701 and 652 potential causal genes were identified respectively, and 297 genes are in common.

Conclusions

The approach we report in this article is highly sensitive and specific. It is not only able to accurately detect SNPs from deeply sequenced exome data, it is also able to detect SNPs by piggybacking on the ChIP-Seq, RNA-Seq and some other sequencing data with low or uneven coverage. We expect our approach to have a wide range of applications.

2. INTRODUCTION

Recent advances of NGS technology reveal limitless insight about the genome, transcriptome, and epigenome of any species. The relatively low cost and incredible throughput of NGS makes it possible to comprehensively catalogue the genetic variation. Projects such as 1000 Genome Project [32] and The Cancer Genome Atlas aim to establish a complete search for variations in common diseases.

SNP calling refers to the determination of the genome positions where there are polymorphisms or at least one of the bases is different from the reference genome. The high error rates in NGS often cause considerable uncertainty for the SNP calling results. Especially, when the coverage is low (< 5 × per site per individual on average), the SNP calling is difficult. To reduce the uncertainty of SNP calling, one proven method is to sequence the targeted region deeply (> 20 × coverage). However, the most common and cost-effective sequencing method is to sequence samples in medium (5-20 × coverage) or low coverage. For example, the 1000 Genome Project sequenced 176 individuals genome-widely at about 3 × coverage, because this design is more efficient to identify rare variants, compared with the design that sequences fewer individuals deeply [59]. Therefore, it is crucial to effectively call SNPs in low coverage data, because the inferred SNPs will influence downstream analysis.

The SNP calling methods can be classified into two types - simple cutoff framework and probabilistic framework. In early studies, the SNP calling analysis first filter the sequencing data, and only the high-confidence bases would be kept. The most common way of filtering the data is to set the threshold of the quality score to Q20.

Bases with quality scores smaller than the threshold are ignored. Then the number of times that each allele is observed will be counted, and the allele with largest count but different from reference genome will be called as SNP. This type of method works well when the sequencing depth is high. Some commercially available softwares such as Roche's GSMapper, the CLC Genomic Workbech software, and the DNSTAR Lasergene software are based on this design. However, for moderate or low coverage sequencing data, this method loses valuable information and results in under-calling. Also, this type of method does not provide a measure for the uncertainty of the inference. Thus, several probabilistic-based methods were developed [60-65] to solve this problem. Briefly, these methods use Bayes' formula to compute the posterior probability for each genotype, and the one with highest posterior probability is generally called. Then SNP is called based on the genotype calling result.

However, most of the existing SNP calling methods do not work well for the sequencing data with low or uneven coverage. Here, we propose a simple Bayesian approach for detecting SNPs, which is highly sensitive and specific, especially for the low coverage sequencing data. We applied our method to a dataset from prostate cancer cell line PC-3, which was originally obtained for epigenetic studies of histone modifications using ChIP-Seq and thus has low and uneven coverage. Compared with Varscan, a popular SNP calling tool that is adaptive to extreme coverage, our method called more SNPs with higher quality. We also applied our method to two exome sequencing datasets from colon cancer cell lines RKO and SW48 respectively, and it shows that our method outperforms both Varscan and DNAnexus - a commercially available SNP calling software. The genome-wide mutations of PC-3, RKO and SW48

had not been studied previously, so our results will provide valuable information for the future cancer study. Also, there are thousands of ChIP-Seq and RNA-Seq experiments conducted every year, and our method can piggyback these data for the SNP analysis. We expect our method to have a wide range of application.

3. MATIRIALS & METHODS

3.1. Data

We applied our method to identify the SNPs of three cancer cell lines – prostate cancer cell line PC-3, colon cancer cell lines RKO and SW48. PC-3 data was generated by Dr. Jean-Pierre Issa's lab and sequenced with Illumina. The data was from 40 lanes including input and Chip-seq. The details of the data source are listed in **Table S1**. Totally about 109M reads with length of 30 or 36 passed the purity filter and contained no ambiguous nucleotide such as "N", and they were used to identify the SNPs in PC3. RKO and SW48 data were from exome capture and RNA-Seq generated by Dr. Marcos Estecio. The reads are 75bp long. 64M reads for RKO and 69M reads for SW48 were used in our analysis.

3.2. Base-calling error probability calculation

Illumina pipeline encodes the quality score from 0 to 62 using ASCII 64 to 126, although only 0 - 40 could be expected in the real data. The quality score is in Phred

format, which is related to the base-calling error probability as shown in the following equation:

$$Q = -10\ \log_{10} p \qquad (\boldsymbol{Eq.3.1})$$

where Q is the quality score and p is the bass-calling error probability. We calculated the accuracy of the quality scores using the PhiX data from 8 independent experiments. **Figure 3.1** shows that the reported quality score is almost identical to the empirical quality score. Therefore, in our analysis, we used the reported quality score by Illumina to calculate the base-calling error probability.

3.3. Cross-talk matrix generation

For a miscalled base, the true nucleotide type is R, the probability of being called to nucleotide type S is defined as cross-talk probability $\alpha_{R,S}$, so $\alpha_{R,S} = \Pr(S|\, miscalled, true\ type\ is\ R)$, where $R, S \in \{A, T, G, C\}$, and $R \neq S$. In this article, we define the cross-talk matrix as the matrix composed of cross-talk probabilities shown as follow:

	A	T	G	C
A	N/A	$\alpha_{A,T}$	$\alpha_{A,G}$	$\alpha_{A,C}$
T	$\alpha_{T,A}$	N/A	$\alpha_{T,G}$	$\alpha_{T,C}$
G	$\alpha_{G,A}$	$\alpha_{G,T}$	N/A	$\alpha_{G,C}$
C	$\alpha_{C,A}$	$\alpha_{C,T}$	$\alpha_{C,G}$	N/A

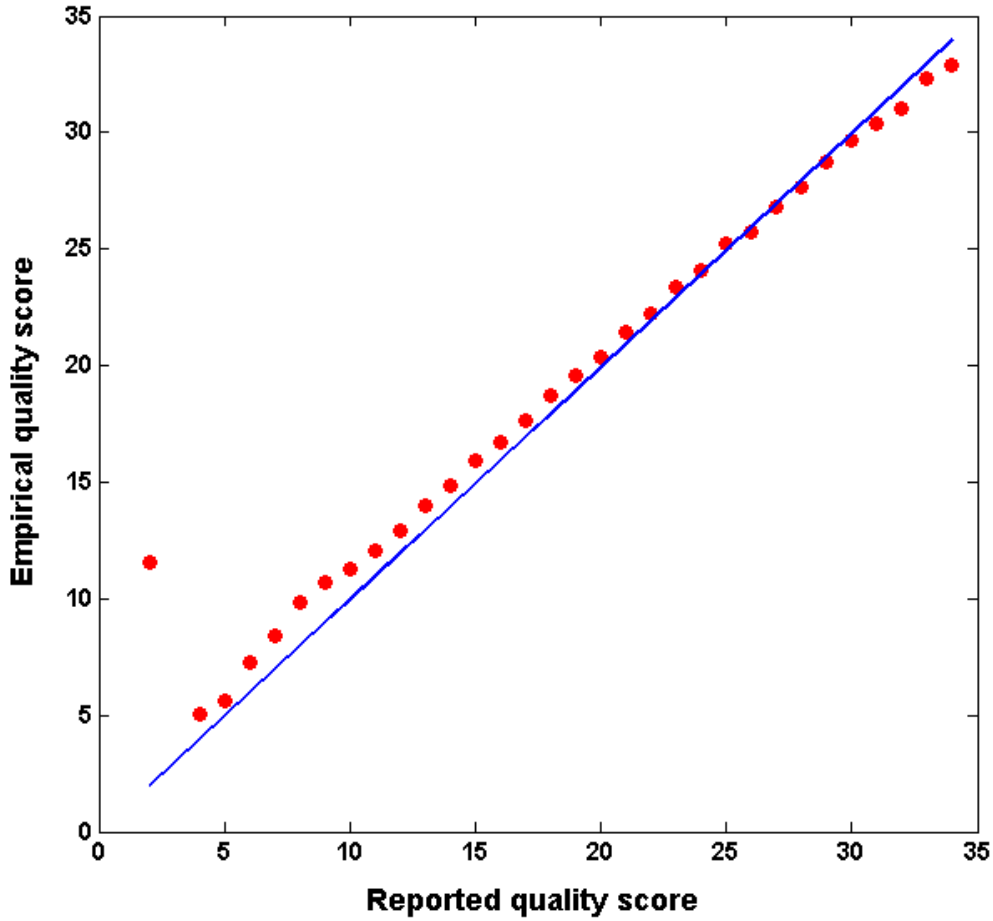


Figure 3.1 plot of reported quality score versus empirical quality score from the PhiX data of 8 independent sequencing experiments. Empirical base-calling error probability for a specified reported quality score was calculated by $p_q = \frac{M_q}{T_q}$, where p_q is the base-calling error probability for reported quality score q, M_q is the count of mismatched bases with reported quality score of q, and T_q is the total count of bases with reported quality score of q. Then the empirical quality score could be obtained using Eq. 3.1.

So $\sum_{S \in \{A,T,G,C\}, S \neq R} \alpha_{R,S} = 1$. We have proved that the cross-talk matrices for every individual lane are consistent in a sequencing run. So we calculate the cross-talk matrix for a sequenced sample by using the corresponding PhiX data in the same run. Denote $c_{R,S}$ the count of bases miscalled from R to S, we calculated $\alpha_{R,S}$ with the following equation:

$$\alpha_{R,S} = \frac{c_{R,S}}{\sum_{I \in \{A,T,G,C\}, I \neq R} c_{R,I}} \qquad (\boldsymbol{Eq.3.2})$$

3.4. Model for SNP calling

A Bayesian algorithm is applied for our SNP discovery. Basically, for a specific genomic position, assuming that we do not know the reference genome, we calculate the probability for each nucleotide type based on the data at this position, and the one with largest probability is the true nucleotide. If this nucleotide is different from the reference genome, we call it a SNP. For a given genomic position, the aligned bases are denoted as *D*. We divide *D* into groups based on their nucleotide types. For example, if the aligned bases are composed of three types of nucleotide – A, G and C, then D can be divided into 3 groups – groups with nucleotide type A, G, or C respectively. Suppose there are *T* groups and the t[th] group has K_t bases, then we have $D = \{D_t\}_{t=1}^{T}$ and $D_t = \{D_{kt}\}_{kt=1}^{K_t}$. Given a nucleotide type N_j, where $N_j \in \{A, C, G, T\}$, we use the following equation to estimate the probability whether N_j is the true nucleotide type.

$$\Pr(N_j|\mathrm{D}) = \frac{\Pr(N_j, \mathrm{D})}{\Pr(D)}$$

$$= \frac{\Pr(N_j, \mathrm{D})}{\sum_{N_i \in \{A,C,G,T\}} \Pr(N_i, \mathrm{D})}$$

$$= \frac{\Pr(D|N_j)\Pr(N_j)}{\sum_{N_i \in \{A,C,G,T\}} \Pr(D|N_i)\Pr(N_i)}$$

Since the reference genome is assumed to be unknown, we give equal prior probability to each nucleotide type, which is $\Pr(A) = \Pr(T) = \Pr(G) = \Pr(C) = 0.25$. Then the above equation is simplified to

$$\Pr(N_j|\mathrm{D}) = \frac{\Pr(D|N_j)}{\sum_{N_i \in \{A,C,G,T\}} \Pr(D|N_i)} \qquad (\boldsymbol{Eq.\,3.3})$$

To estimate $\Pr(D|N_j)$ in **Eq. 3.3**, we define C_t as the nucleotide type for t[th] group. For base D_{kt}, $\{e_{kt,N_j} = 0\}$ and $\{e_{kt,N_j} = 1\}$ respectively denote match or mismatch between D_{kt}'s nucleotide type C_t and the conditional nucleotide type N_j. Thus, we have

$$\Pr(D|N_j) = \prod_{t=1}^{T} \Pr(D_t|N_j)$$

$$= \prod_{t=1}^{T} \prod_{kt=1}^{K_t} (1 - p_{kt})^{1-e_{kt,N_j}} (p_{kt}\alpha_{N_j,C_t})^{e_{kt,N_j}} \qquad (\boldsymbol{Eq.\,3.4})$$

By pluging **Eq. 3.4** into **Eq. 3.3**, the probability that N_j is the true base is

$$\Pr(N_j|\mathrm{D}) = \frac{\prod_{t=1}^{T} \prod_{kt=1}^{K_t} (1 - p_{kt})^{1-e_{kt,N_j}} (p_{kt}\alpha_{N_j,C_t})^{e_{kt,N_j}}}{\sum_{N_i \in \{A,C,G,T\}} \prod_{t=1}^{T} \prod_{kt=1}^{K_t} (1 - p_{kt})^{1-e_{kt,N_i}} (p_{kt}\alpha_{N_i,C_t})^{e_{kt,N_i}}} \qquad (\boldsymbol{Eq.\,3.5})$$

Where p_{kt} is the base-calling error probability for base D_{kt}, and can be obtained from **Eq. 3.1**. $\alpha_{N_j,C_t}$is the cross-talk probability that is calculated from **Eq. 3.2**. With **Eq. 3.5**, we calculate the conditional probability for each nucleotide type. The one with maximal conditional probability is the nucleotide we detect at this position, and its corresponding probability is denoted with β. If the detected nucleotide is different from reference genome, we define it as a SNP.

4. RESULTS

4.1. Cross-talk study

Cross-talk is one of the major sources of error for Illumina sequencing. The Illumina Genome Analyzer uses two lasers and four filters to detect the nucleotides labeled with different dyes. Due to the overlap of the emission spectra of the four fluorophores, the detected images are not independent. The intensities of A and C are correlated as are those of G and T [66, 67]. Thus, the probabilities of miscalling for different nucleotide types should vary.

. To control the quality and facilitate base calling, PhiX, a virus with a small and well-defined genome containing about 45% GC and 55% AT, is sequenced together with sequencing samples. Because of its properties, PhiX sequencing data is suitable for the cross-talk study. Firstly, we compared the cross-talk of PhiX data from 8 independent runs. To evaluate the cross-talk, we divided the sequencing errors into

groups based on the nucleotide types of the true base and the miscalling base, and calculated the sequencing error rates for each group by

$$e_{R,S} = \frac{count\ of\ bases\ miscalled\ from\ R\ to\ S}{total\ number\ of\ mismatches}$$

Where *R* is the nucleotide on the reference genome, and *S* is the aligned nucleotide from the sequencing reads. For example, for $e_{A,C}$ the numerator is the number of Cs in the reads that are aligned to As in the reference genome, and the denominator is the total number mismatched bases in the reads. Because $R \neq S$, there are totally 12 types of $e_{R,S}$ and their summation should be 1. As shown in **Figure 3.2,** the cross-talk patterns for different runs are not consistent. For instance, $e_{T,C}$ and $e_{G,C}$ are very large for data 5, but very small for data 4. Secondly, we compared the cross-talk among different lanes in a single run. Since PhiX is sequenced for calibration purpose, usually it only takes one lane in a run. Thus, additionally to PhiX data, we used two lanes of reads sequenced from *E.coli* to study the cross-talk among different lanes. *E.coli* genome is relatively small and contains far fewer SNPs compared with Mammalian, so it is suitable for the sequencing error study. In **Figure 3.3 (a)**, Lane 1 and Lane 2 are from *E.coli* sequencing reads, while Lane 8 is from PhiX sequencing reads in the same run. Obviously, these three lanes share the same cross-talk pattern. Lastly, we looked at the cross-talk for different tiles in the same lane. A lane of PhiX sequencing reads were divided into two groups: one composed of 1-49 tiles, and one composed of 50-100 tiles. We found that the sequencing patterns of these two groups are quite similar (**Figure 3.3 (b)**). Therefore, the cross-talk patterns vary for different sequencing runs, but are consistent in each single run.

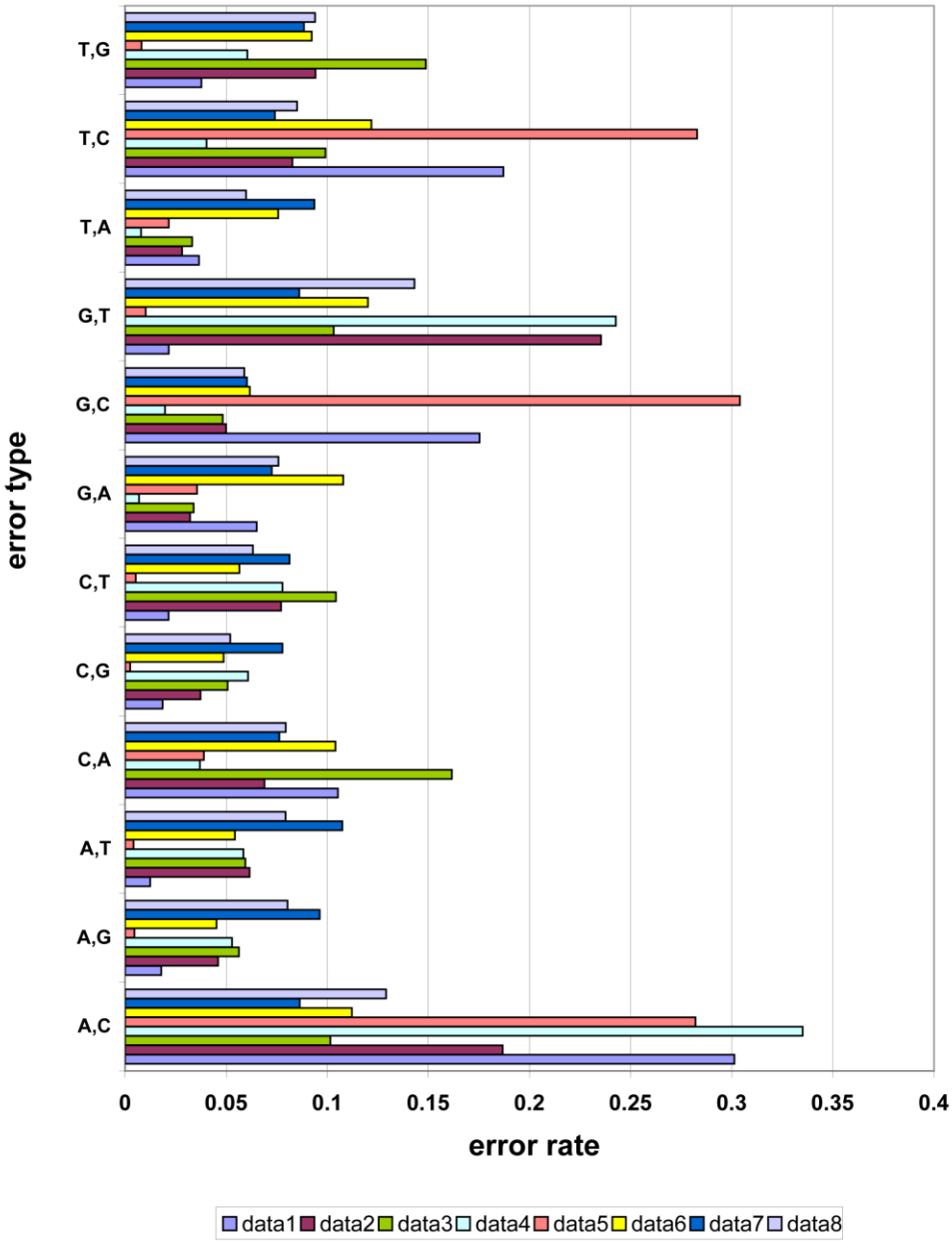


Figure 3.2. Cross-talk patterns for different runs are not consistent. Sequencing errors were classified for 8 sets of PhiX data from 8 independent sequencing runs. The vertical axis is the error type based on the nucleotide type. For example, "A,C" represents that the true nucleotide is A, but it is miscalled to C. The horizontal axis is the relative error rate, which is calculated by $e_{R,S} = \frac{count\ of\ bases\ miscalled\ from\ R\ to\ S}{total\ number\ of\ mismatches}$.

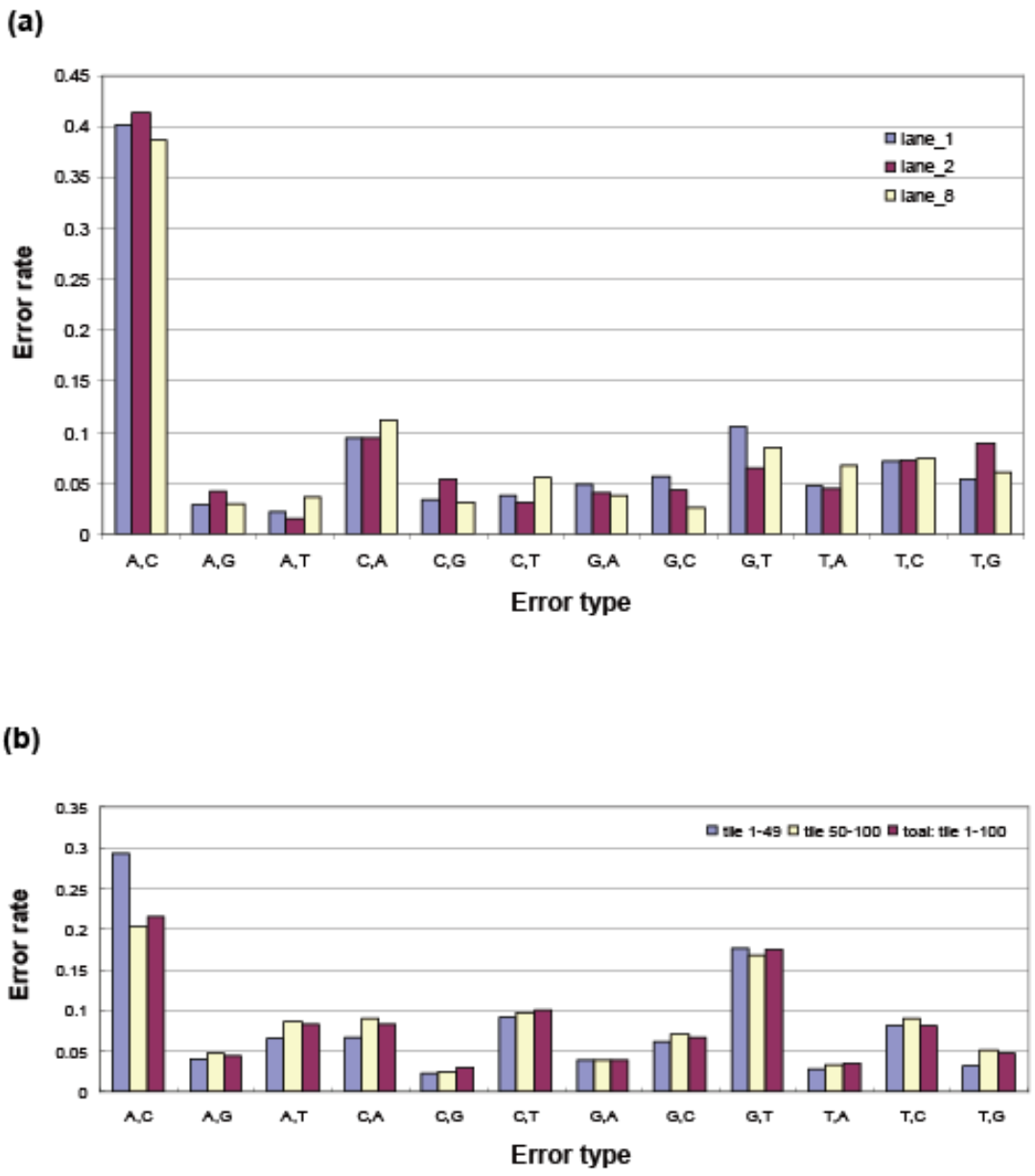


Figure 3.3. Cross-talk patterns are consistent in a run. Sequencing errors were classified in the same way as Figure 3.2. (a) We compared the cross-talk patterns of different lanes from the same sequencing run. Lane 1 and Lane 2 were sequenced from *E.coli* genome. Lane 8 is sequenced from PhiX. (b) A single lane of PhiX sequencing data was divided into two groups based on tiles. The cross-talk patterns of these two groups are consistent.

4.1. Model testing with PhiX data

We tested our method with the PhiX data by manually mutating some positions on the PhiX genome, and using the mutated genome as the reference genome. Thus, the mutated positions serve as the gold standard of SNP sites. As shown in **Figure 3.4**, the coverage of the PhiX genome by the sequencing reads is high and uniform. The average sequencing depth is about 73. We applied our method to this dataset, and compared our results with the gold standard mutations (**Figure 3.5**). The SNPs called by our method exactly match the gold standard. All the SNPs were correctly called, and no false positive SNPs were observed.

4.1. SNPs detection for prostate cancer cell line PC-3

Prostate cancer usually occurs in old men, and is the most common cause of death from cancer in men over age 75. PC-3 is a human prostatic carcinoma cell line, which is initiated from a bone metastasis of a grade IV prostatic adenocarcinoma from a 62-year-old man [68, 69]. PC-3 is near-triploid with a modal number of 62 chromosomes. PC-3 has a unique karyotype. Normal chromosomes N2, N3, N4, N5, N12, and N15 are absent. The mutation status of PC-3 is still unknown. The PC-3 data we obtained were originally generated for the epigenetic studies of histone modification with ChIP-Seq, so its coverage is low and uneven. Therefore, PC-3 dataset is perfect for testing the performance of our method on low coverage data.

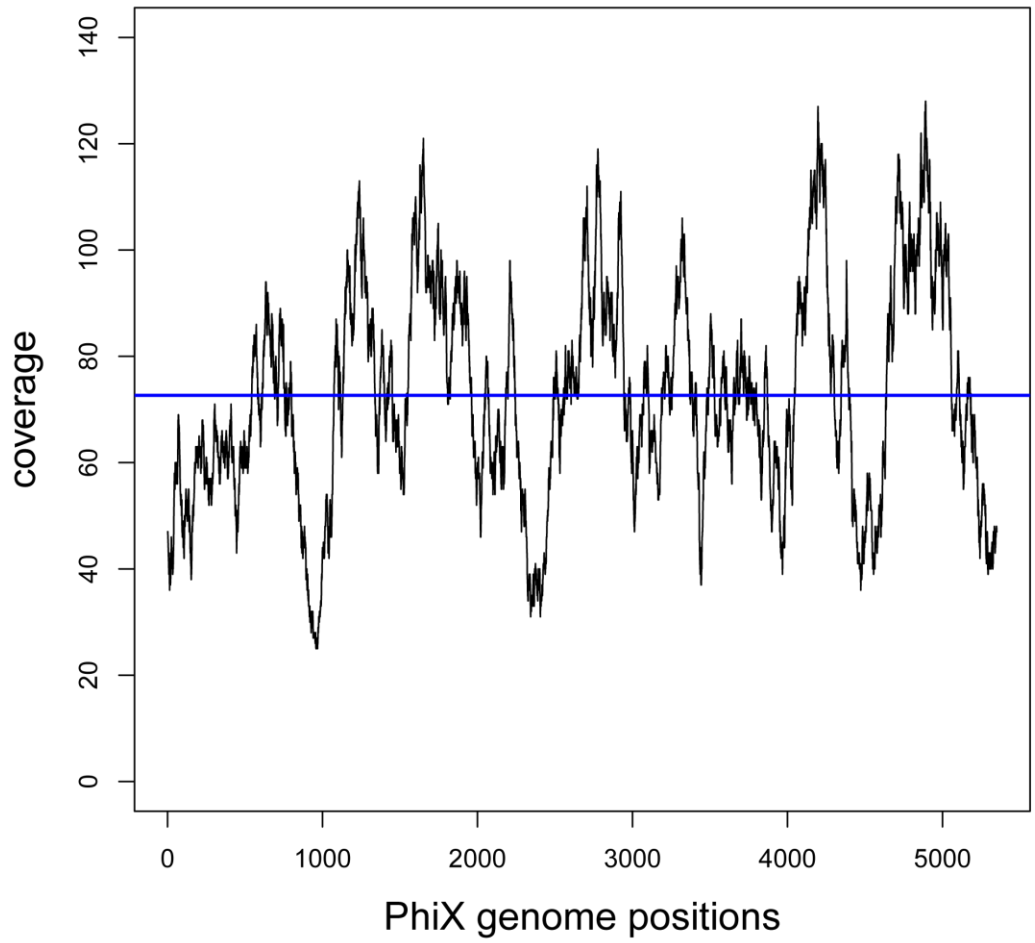


Figure 3.4. Genome coverage of PhiX by the sequencing reads used for testing our method. The coverage is high and uniform. Average coverage is about 73, which is shown as the blue line.

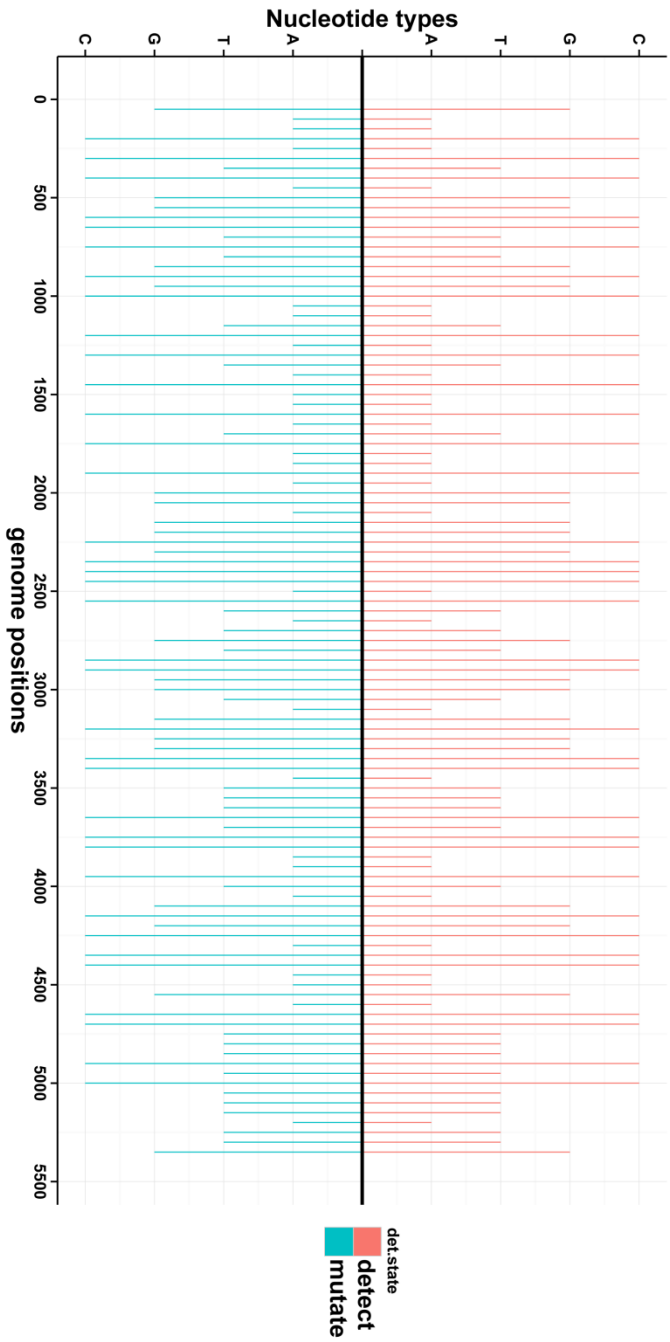


Figure 3.5 Comparison of the gold standard mutations and the SNPs detected by our method. The cyan color represents the gold standard mutations we generated, and the orange color represents the SNPs we detected. The mutations we generated are identical to the mutations detected by our methods.

4.2.1. Coverage of PC-3 sequencing data

We aligned the reads of PC-3 dataset with Bowtie's v-mode by allowing up to 3 mismatches for the whole read [56], and calculated the coverage for different regions in human genome with BEDTools [70]. There are many gaps in the reference human genome, which cannot be sequenced or mapped. The largest gaps exist in the highly repetitive regions of the genome - mostly around the centromeres and other heterochromatic regions. Some gaps also locate at gene clusters. To obtain an accurate coverage, we subtracted these gaps from the reference genome and used the gapless genome in the calculation. As shown in **Table 3.1**, ~50% of the genome is sequenced, and most of the sequencing reads aligned to the non-exonic regions. ~ 90% of the sequenced positions are covered with 4 or fewer reads, and only < 5% of the sequenced positions are covered with 6 or more reads (**Figure 3.6**). The average sequencing depths for different regions are less than 3, while the maximum sequencing depths are very large, indicating that the coverage is low and uneven. The average sequencing depth here is the average from the covered positions.

4.2.1. Characterization of the identified SNPs for PC-3

To assess the specificity of our method, we compared the identified SNPs with dbSNP (Single Nucleotide Polymorphism Database) and 1000-genome database. dbSNP is a free public archive for genetic variation developed and hosted by NCBI. Until now, most variants have been catalogued by dbSNP. Since most genetic variants in one individual should have been previously observed from other people, usually

Table 3.1 Coverage of PC-3 dataset

	Sequencing depth		Covered regions	
	Maximum	**Average**	**Number of bases (Mb)**	**Percentage**
Whole-genome	12429	1.2	1419.7	49.60%
Gene region	771	1.3	1276.3	52.70%
Exome	771	2.6	83.2	66.30%
CDS (coding sequence)	348	2.6	41.7	72.80%

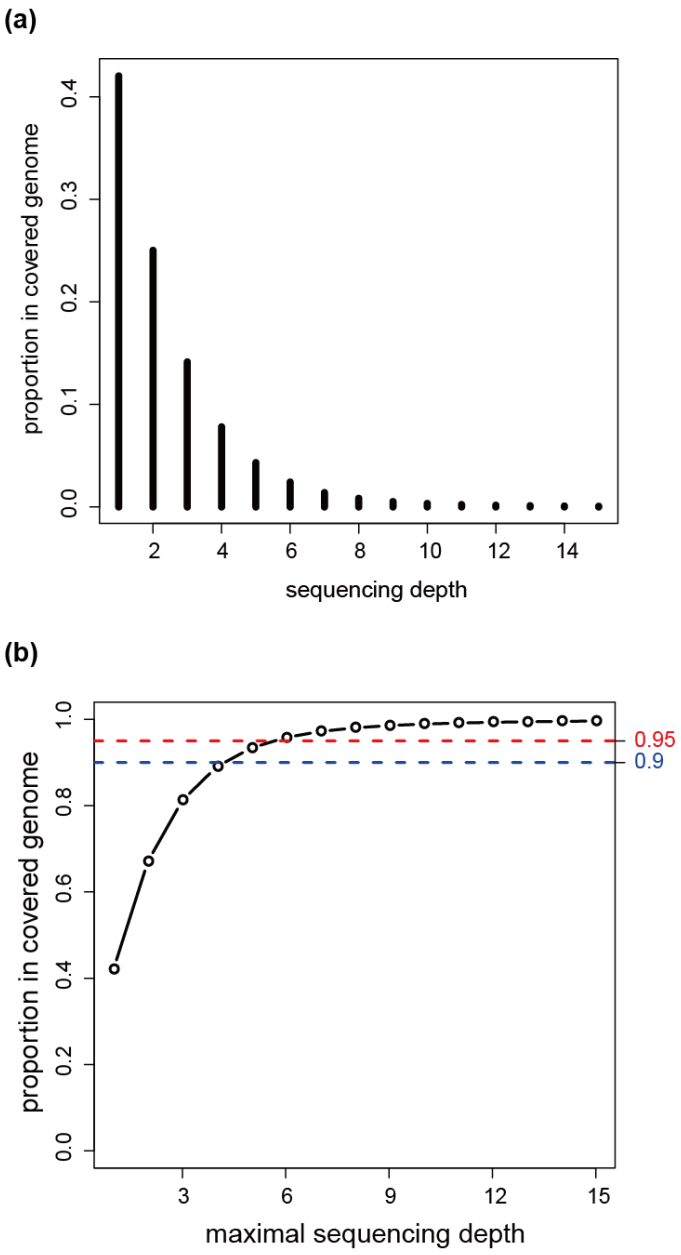


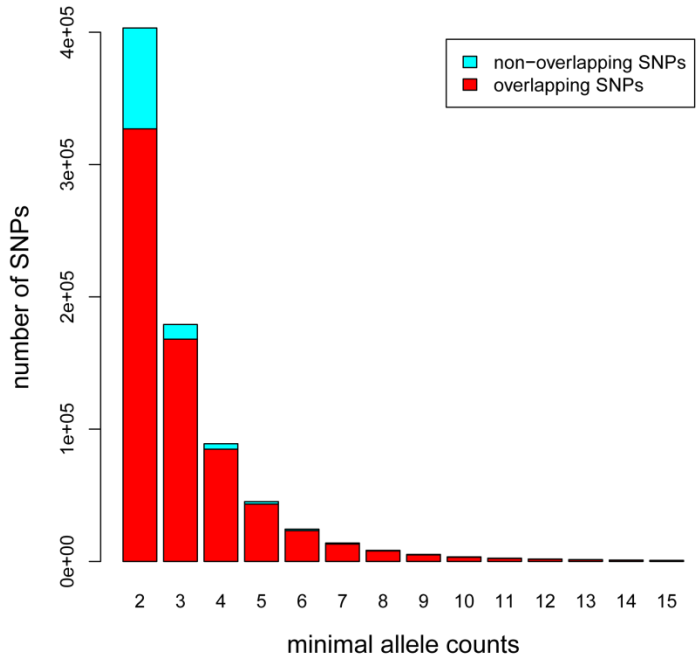
Figure 3.6 The distribution of the sequencing depth of PC-3 dataset in the whole genome. (a) Density distribution: the proportion of covered positions at different sequencing depth. (b) Cumulative distribution: proportion of the positions covered at equal or less than specified depth. So the sequencing depth is very low, and about 95% of the sequenced genome is covered with 6 or fewer reads. Note that the proportion here is calculated with respect to the positions with at least one read aligned.

dbSNP is used as the standard to measure the specificity of SNP detection. dbSNP has been constantly updated, so >90% of the SNPs are expected to be discovered in dbSNP . Although the presence in dbSNP does not absolutely confirm the authenticity of the SNP, since the dbSNP build 135 contains 47.8 million SNPs (only 1.6% of the whole genome) , the relative difference between call sets should be able to reflect the quality differences. Similarly to dbSNP, 1000-genome database is also a large public variation data resource, which is established by sequencing the genomes of a large number of people. The current 1000-genome database contains 38 million SNPs (1.3% of the whole genome). In this article, we use dbSNP rate as the percentage of the SNPs discovered in dbSNP, and use 1000-genome rate as the proportion of the SNPs described in 1000-genome database.

The majority of the SNPs in PC-3 detected by our method are previously described by dbSNP and 1000-genome database, suggesting that our method works very well on PC-3 data. As shown in **Figure 3.7(a) and Figure 3.8(a)**, the dbSNP rate and 1000-genome rate increase when the cutoff of the SNP allele count increases. For the SNPs with allele counts equal or larger than 3, over 90% of the SNPs are discovered in the dbSNP or 1000-genome database. Recall that β denotes the conditional probability of called SNP. With the increase of β value, the dbSNP rate and 1000-genome rate also increase (**Figure 3.7(b) and Figure 3.8(b)**).

To improve the SNP calls, we use the product of allele count and β to control the quality. SNPs with $allele\_count * \beta$ below the cutoff value will be removed from the final result. Setting the cutoff too large will miss a lot of true SNPs, while setting the cutoff too small will produce many false positive SNPs. Thus, varying the value of this

(a)



(b)

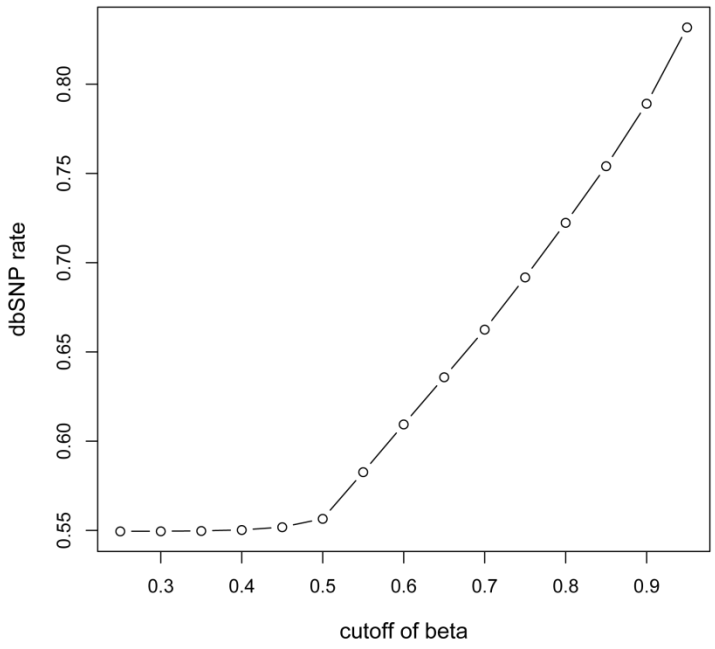
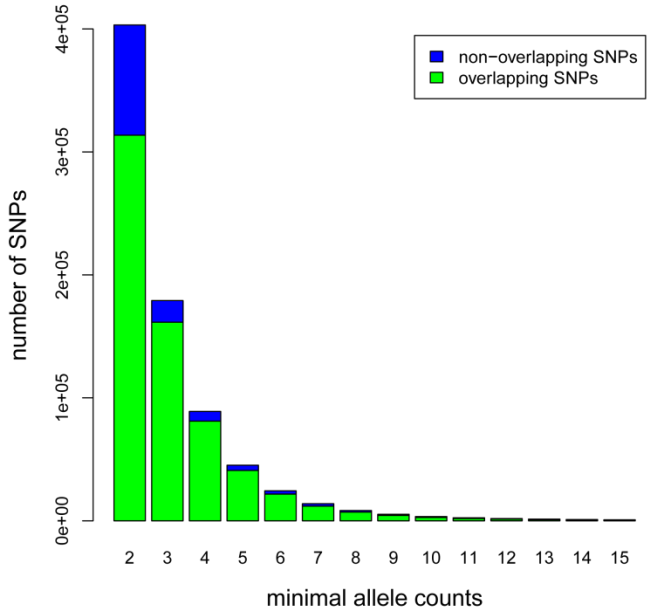


Figure 3.7 Comparison of PC-3 SNPs detected with our method to dbSNP. With the increase of cutoff for (a) SNP allele count and (b) β value, dbSNP rate increases.

(a)



(b)

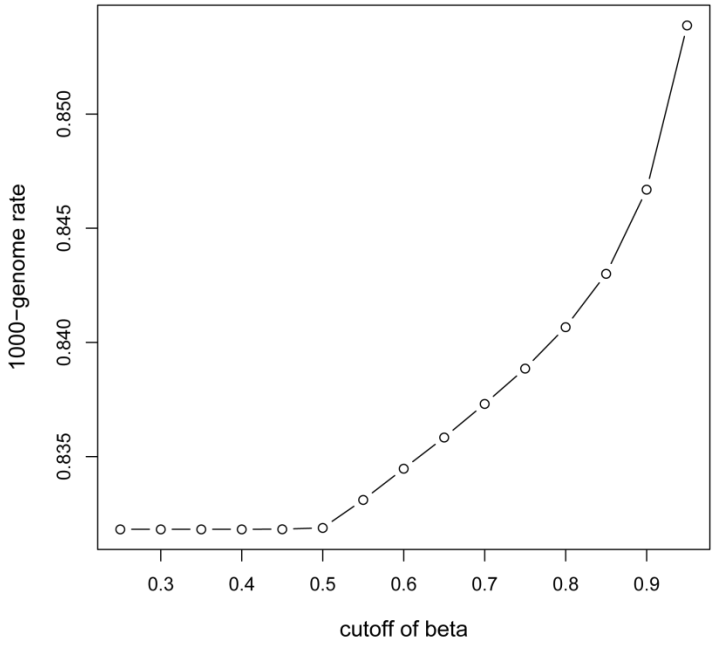


Figure 3.8 Comparison of PC-3 SNPs detected with our method to 1000-genome database. With the increase of cutoff for (a) SNP allele count and (b) β value, 1000-genome rate increases.

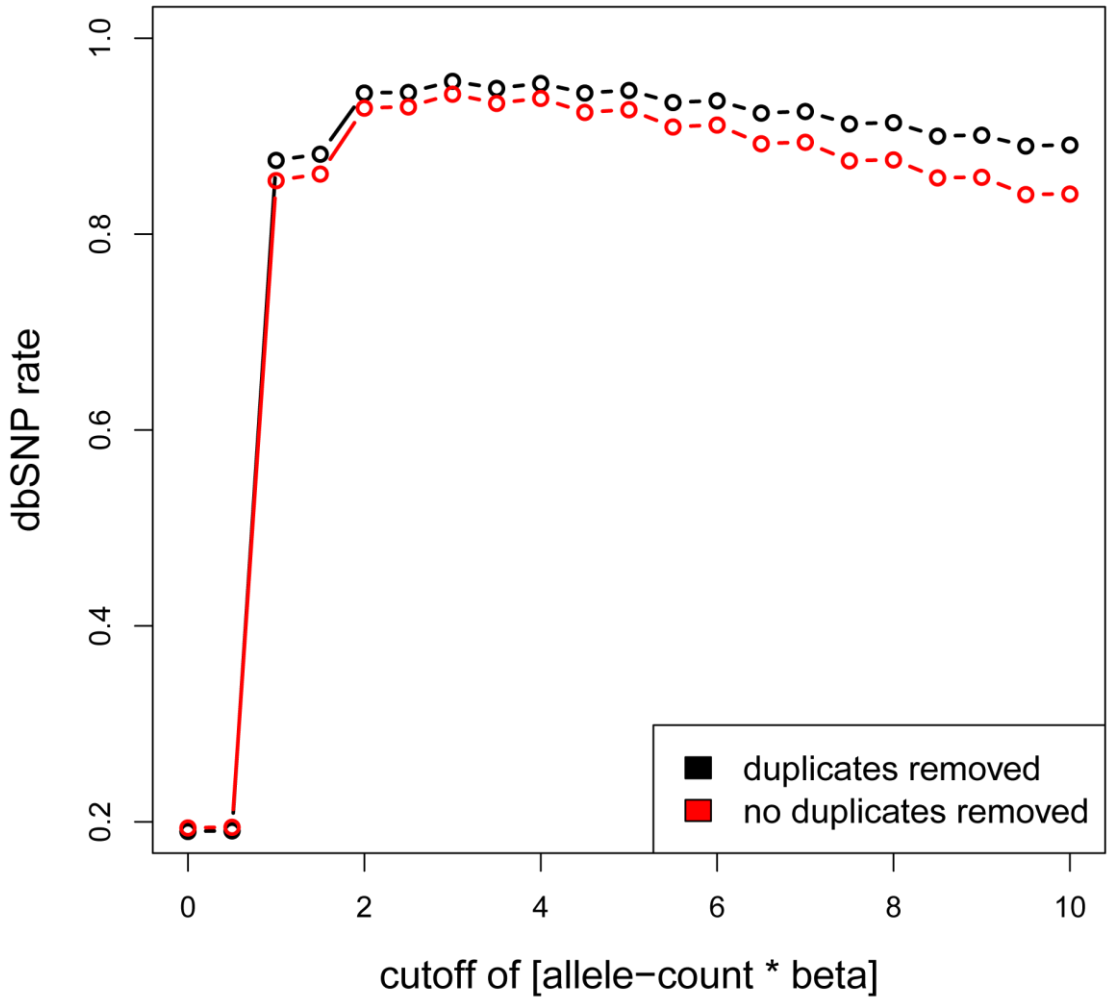


Figure 3.9 Removing the Duplicates slightly increase the dbSNP rate. At various cutoff point of allele-count * β, the dbSNP rates after removing the duplicates become higher.

cutoff value results in a trade-off between sensitivity and specificity. For PC-3 data, to balance the sensitivity and specificity, we set the cutoff value to 2, with which about 94.4% of SNPs overlap with dbSNP (**Figure 3.9**).

Duplicate reads contribute to the false positive SNPs. When the sequencing library preparation involved the PCR amplification step, duplicated reads are usually observed. Duplicate reads from PCR amplification usually results in areas of high disproportional high coverage, and are often the cause of false positive in SNP calling, especially when the replication errors are made by the enzymes during the amplification. So we removed the duplicate reads that share common coordinates, sequencing direction and same sequence. The effect of removing duplicates varies in different datasets. For PC-3, removing the duplicates slightly improved the SNP calling, and the improvement becomes more significant when the cutoff value increases (**Figure 3.9**). It is reasonable because the PC-3 data is from many experiments and the majority of the PC-3 covered positions has low sequencing depth, indicating that the duplicate rate is low. The duplicate rate is larger for the positions with higher sequencing depth, so the effect of removing duplicates is more significant for higher depth.

4.2.2. Comparison with other SNP detection tools

To evaluate the performance of our method, especially its ability of detecting SNPs from low coverage data, we compared our method with Varscan - a very popular variant-calling method, which has the adaptability to extreme read depth and pooled samples. It employs a robust heuristic/statistic approach to call variants. With optimized

setting for the parameters, Varscan called 115,571 SNPs, of which 101,736 are observed in dbSNP. Our method detected 526,925 SNPs and 497,526 of them are discovered in dbSNP (**Table 3.2**). So our method found about 4.5 times more SNPs than Varscan, while maintaining a higher dbSNP rate (94.42% versus 88.03%), indicating that our method is more sensitive and specific than Varscan for PC-3 data.

Transition/transversion ratio (Ti/Tv) is a critical metric to assess the specificity of SNP-calling. Transition is the substitution of a purine by a purine or a pyrimidine by a pyrimidine ($T \leftrightarrow C, A \leftrightarrow G$). Transversion is a change from purine to pyrimidine, or vice versa ($T \leftrightarrow A, T \leftrightarrow G, C \leftrightarrow A, C \leftrightarrow G$). Ti/Tv is 0.5 when there is no bias towards either transition or transversion, because the two kinds substitutions are equal probable, and there are twice as many transversions than transitions. However, for all the genomes examined so far, transitions occur more frequently than transversions [73-77]. Ti/Tv is known to be a general property of DNA sequence evolution. Inter-species comparison and previous sequencing projects showed that the Ti/Tv for the genome-wide variants is ~2.0-2.1, and the exonic Ti/Tv is ~3.0-3.5 [78, 79]. Given the observed Ti/Tv, the false discovery rate (FDR) can be obtained by

$$FDR = 1 - \frac{Ti/Tv_{observed} - 0.5}{Ti/Tv_{expected} - 0.5}$$

where the $Ti/Tv_{expected}$ is the expected value of Ti/Tv. We let $Ti/Tv_{expected}$ equal 2.05 for whole genome and 3.25 for exome. As shown in **Table 3.2**, Ti/Tv obtained by our method is 1.98, which is more close to the standard genome-wide Ti/Tv (2.0-2.1) than the Varscan Ti/Tv (1.68). Also, the FDR of our method is only 0.04, which is <0.05 and much better than the Varscan FDR (0.24).

Table 3.2 Comparison of our method with Varscan for PC-3 data. In the calculation of FDR, 2.05 is used as the expected Ti/Tv ratio.

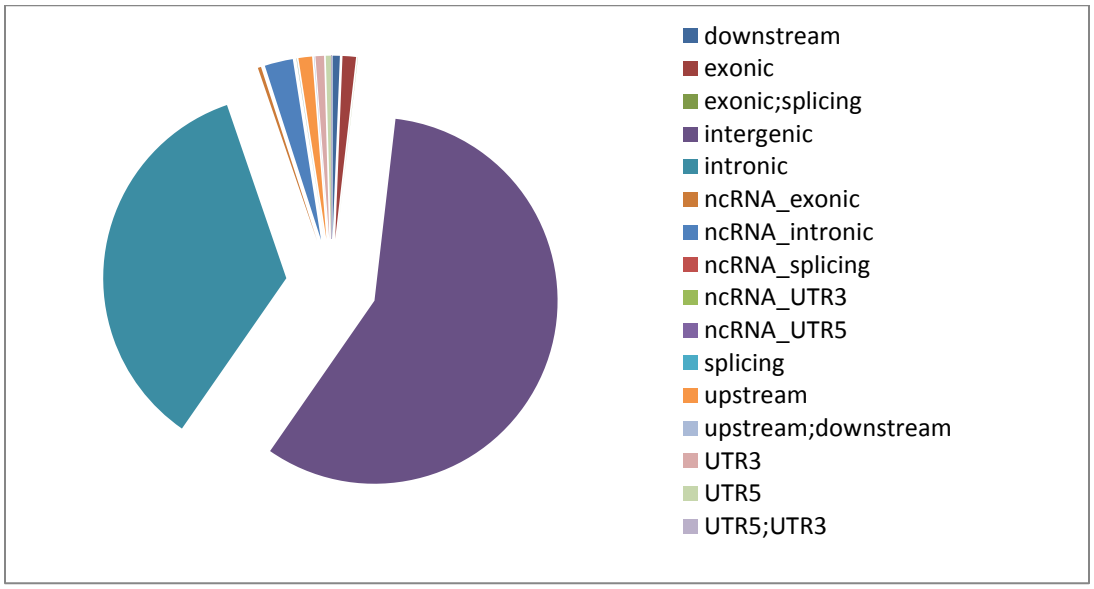
	our method	**Varscan**
Number of called SNPs	526,925	115,571
Number of SNPs observed in dbSNP	497,526	101,736
dbSNP rate (%)	94.42	88.03
Ti/Tv ratio	1.98	1.68
FDR	0.04	0.24

4.2.3. Annotation of SNPs in PC-3

We annotated the SNPs with ANNOVAR - a tool for functionally annotating the genetic variants [80]. Based on the function of the DNA sequences where the SNPs aligned, SNPs are classified as intergenic SNPs, intronic SNPs, exonic SNPs, splicing site SNPs, upstream/downstream SNPs, 5' / 3' UTR (untranslated region) SNPs, and ncRNA (non-coding RNA) SNPs. Here, splicing site SNP is the SNP within 2-bp of a splicing junction, and upstream/downstream SNP is the SNP overlaps 1-kb region upstream/downstream of transcription start/end site. Consistent to the distribution of the sequencing reads of PC-3, most of the SNPs detected by our method are in intergenic and intronic regions (**Figure 3.10 (a)**). We also found that the dbSNP rates of different regions vary (**Figure 3.10 (b)**). The splicing site SNPs for both RNA and ncRNA have lower dbSNP rates than the other regions. The detailed numbers are listed in **Table 3.3 (a)**. Based on the effect of the substitution to the genetic coding, exonic SNPs are further grouped into synonymous SNPs, nonsynonymous SNPs, stopgain SNPs, stoploss SNPs, and unknown SNPs (**Table 3.3 (b)**). Stopgain SNPs are very important, because they result in truncated, incomplete, and usually nonfunctional protein product. There are four novel stopgain SNPs and one novel stoploss SNP discovered in PC-3 data (**Table S2**).

To identify the potential causal genes responsible for the prostate cancer, we prioritized the SNPs in several steps (**Figure 3.11**). Although exons only constitute about 1% of the human genome [15], it is estimated that the protein coding regions constitute about 85% of the disease-causing variants [81]. There is also highly functional variation in the splicing sites [82], so we first perform a gene-based

(a)



(b)

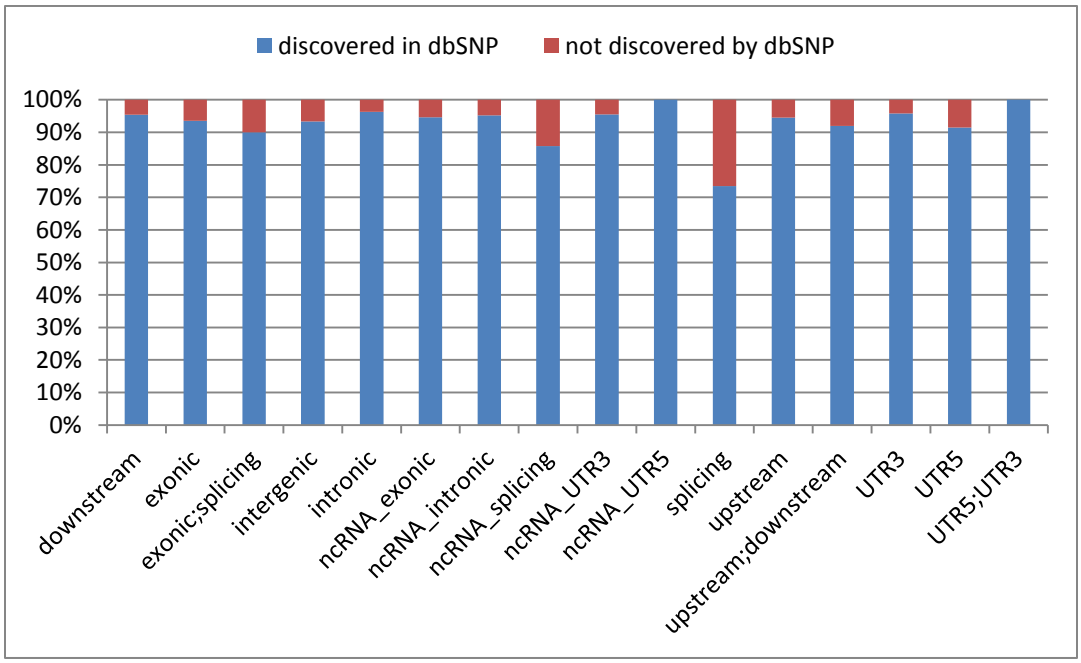


Figure 3.10. (a) Pie chart of PC-3 SNPs for different genome regions. (b) Percentage of SNPs observed in dbSNP for different regions.

Table 3.3 Classification of PC-3 SNPs.

(a) Classification of PC-3 SNPs by genome region.

	Number of called SNPs	Number of SNPs observed in dbSNP	dbSNP rate (%)
downstream	3,361	3,206	95.39
exonic	6,243	5,839	93.53
exonic; splicing	60	54	90.00
intergenic	304,694	284,260	93.29
intronic	184,877	177,945	96.25
ncRNA_exonic	1,505	1,424	94.62
ncRNA_intronic	13,178	12,542	95.17
ncRNA_splicing	7	6	85.71
ncRNA_UTR3	66	63	95.45
ncRNA_UTR5	30	30	100.00
splicing	49	36	73.47
upstream	6,305	5,961	94.54
upstream; downstream	199	183	91.96
UTR3	3,948	3,780	95.74
UTR5	2,400	2,194	91.42
UTR5;UTR3	3	3	100.00

(**b) Classification of exonic SNPs of PC-3 by genetic coding.**

	Number of called SNPs	Number of SNPs observed in dbSNP
nonsynonymous	2,991	2,735
synonymous	3,169	3,029
stopgain	25	21
stoploss	3	2
unknown	115	106

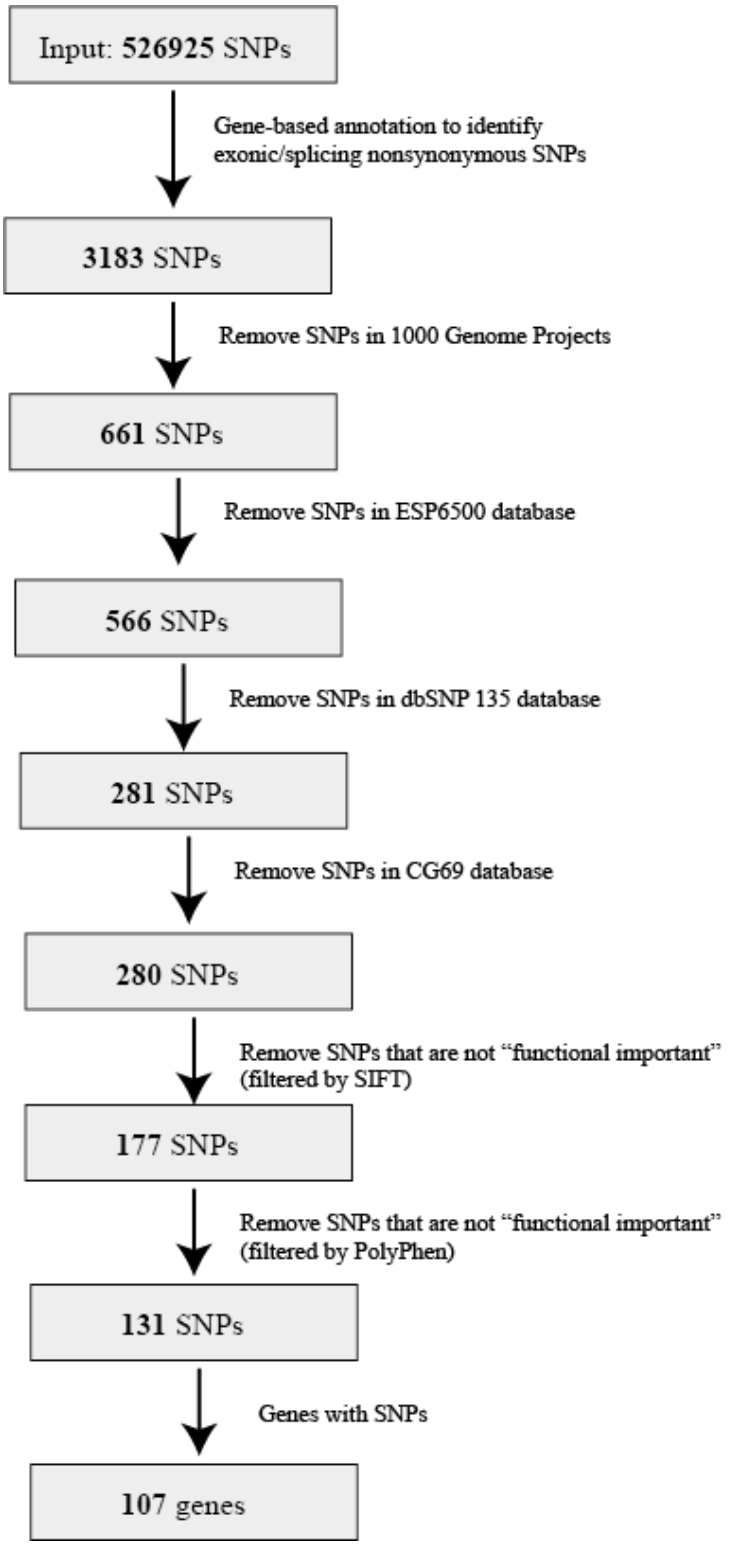


Figure 3.11. Prioritization of causal genes for PC-3.

annotation to identify 526,925 SNPs locating at exons or splicing sites. We then filtered the SNPs from 1000-genome database, ESP (NHLBI GO Exome Sequencing Project) 6500 database, dbSNP135 database and CG (Complete Genomics) 69 database, assuming that the SNPs observed in public databases are less likely to be causal SNPs of cancer. ESP is to discover novel genes and mechanism contributing to the lung, heart and blood disorder, and its database is constituted of the variants sequenced from 6500 exomes of the human genome across diverse, richly-phenotyped populations. CG69 is the variant database established by Complete Genomics Company by sequencing 69 whole human genomes. The novel SNPs were then scored by SIFT [83] and PolyPhen [84] - tools to predicts the importance of variants to the protein function. Generally, SNPs with SIFT score ≤ 0.05 or PolyPhen score ≥ 0.85 are predicted to be deleterious. Totally, 131 SNPs passed all the filters and 107 genes were identified as the causal genes, of which the detail information is listed in **Table S3**.

4.2. SNPs detection for colon cancer cell lines RKO and SW48

American Cancer Society reported that colon cancer is one of the leading causes of cancer-related deaths in the United States. However, colon cancer can often be completely cured with early diagnosis. Identification of the causal variants for colon cancer may greatly contribute to the early diagnosis, and therefore can be particularly significant. RKO and SW48 are two colon cancer cell lines that are used as in vitro models for colorectal cancer to study biochemical mechanisms of carcinoma formation. However, the mutation statuses of RKO and SW48 remain unknown. On the other hand,

exome sequencing is the most popular sequencing strategy for the variants detection. Compared with the whole genome sequencing, exome sequencing is cheaper but still effective, because exome is usually more straightforwardly related to the diseases. Different from whole genome sequencing, exome sequencing data usually has higher coverage and contains more duplicates. Here, we applied our method to detect the SNPs in RKO and SW48 cell lines, not only showing the performance of our method on the exome sequencing data, but also providing valuable information for the future colon cancer study.

4.3.1. Coverage of RKO and SW48 sequencing data

RKO and SW48 are sequenced by exome capture and RNA-seq, so most of the reads align to the exome. The coverage of PKO and SW48 data are similar (**Table 3.4**), because same protocols were applied in the experiments. ~ 60% of exome and ~ 80% of CDS region are sequenced, with the average sequencing depths of ~65 for exome and ~35 for CDS. Compared with PC-3 data, the sequencing depths of RKO and SW48 are much higher and more evenly distributed. ~95% of the sequenced positions are covered by 1-200 reads (**Figure 3.12**).

4.3.1. Characterization of the identified SNPs for RKO and SW48

Similar to PC-3 data, the dbSNP rates and 1000-genome rates of RKO and SW48 also positively relate to the SNP allele count and β value. With an appropriate

Table 3.4 Coverage of RKO and SW48 dataset.

		sequencing depth		covered regions	
		maximum	**average**	**number of bases (Mb)**	**percentage**
RKO	**Exome**	293,247	35.5	76.2	60.67%
	CDS	63,609	64.7	47.2	82.36%
SW48	**Exome**	329,032	34.7	79.2	63.12%
	CDS	86,703	63.2	48.1	83.96%

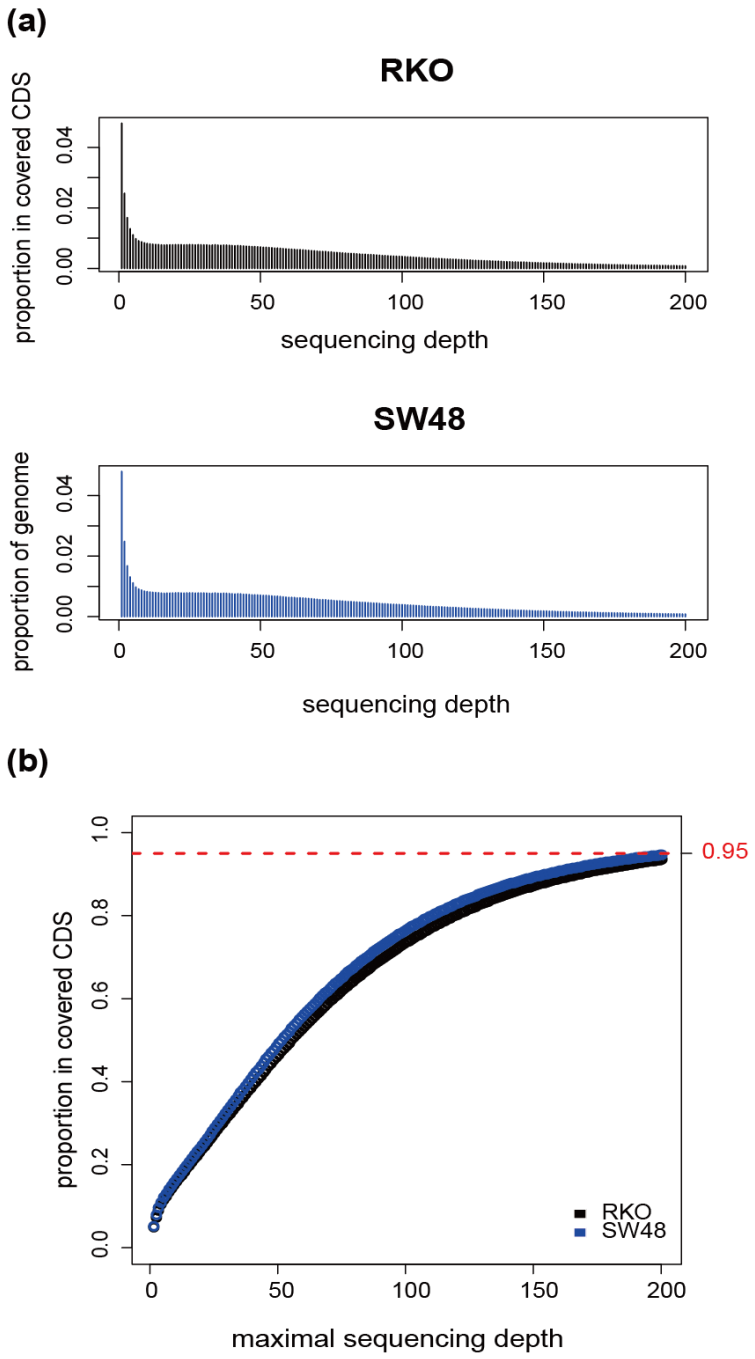


Figure 3.12 The distribution of the sequencing depth of RKO and SW48 datasets in the CDS region. (a) Density distribution: proportion of the covered positions at different sequencing depths. (b) Cumulative distribution: proportion of the positions covered at equal or less than specified depth. Therefore, majority of the sequencing depths almost uniformly distributed in from 1-100, and ~95% of the CDS region has ≤ reads aligned. Note that the proportion here is calculated with respect to the positions in CDS that are covered with at least one read.

cutoff for the product of SNP allele count and β value, we identified 95,142 SNPs for RKO, of which 76,836 are discovered in dbSNP. Similarly, 101,088 SNPs are detected from SW48 dataset, and 76,560 of them are overlapped with dbSNP. The dbSNP rates are 80.76% and 80.90% respectively.

The effect of the duplicate reads was measured. After the duplicate reads are removed, the dbSNP rates show a trend to increase (**Figure 3.13**). Compared with PC-3, the increases for RKO and SW48 are more significant. PC-3 dataset is from many experiments with low sequencing depth, so the duplicate rate is low. RKO and SW48 datasets contain exome sequencing reads which usually have high duplicate rate. Therefore, removing duplicates could highly improve the SNP calling quality for RKO and SW48 datasets.

4.3.2. Comparison with other SNP detection tools

To evaluate the performance of our method on exome capture data, we also ran the RKO and SW48 dataset with Varscan and DNAnexus. DNAnexus is a commercial software for NGS analysis including variation detection. For RKO, Varscan detects 101,088 SNPs with 76,560 overlapped with dbSNP (**Table 3.5**). Compared with Varscan, our method detects ~300 more possibly right variants (variants described in dbSNP) and obtains higher dbSNP rate (80.76% versus 75.74%). Our method also outperforms DNAnexus on SNPs detection. DNAnexus detects much fewer SNPs in dbSNP (52,923), and the dbSNP rate (59.57%) is much lower than our method (80.76%). The Ti/Tv ratios of Varscan and our method are higher than that of

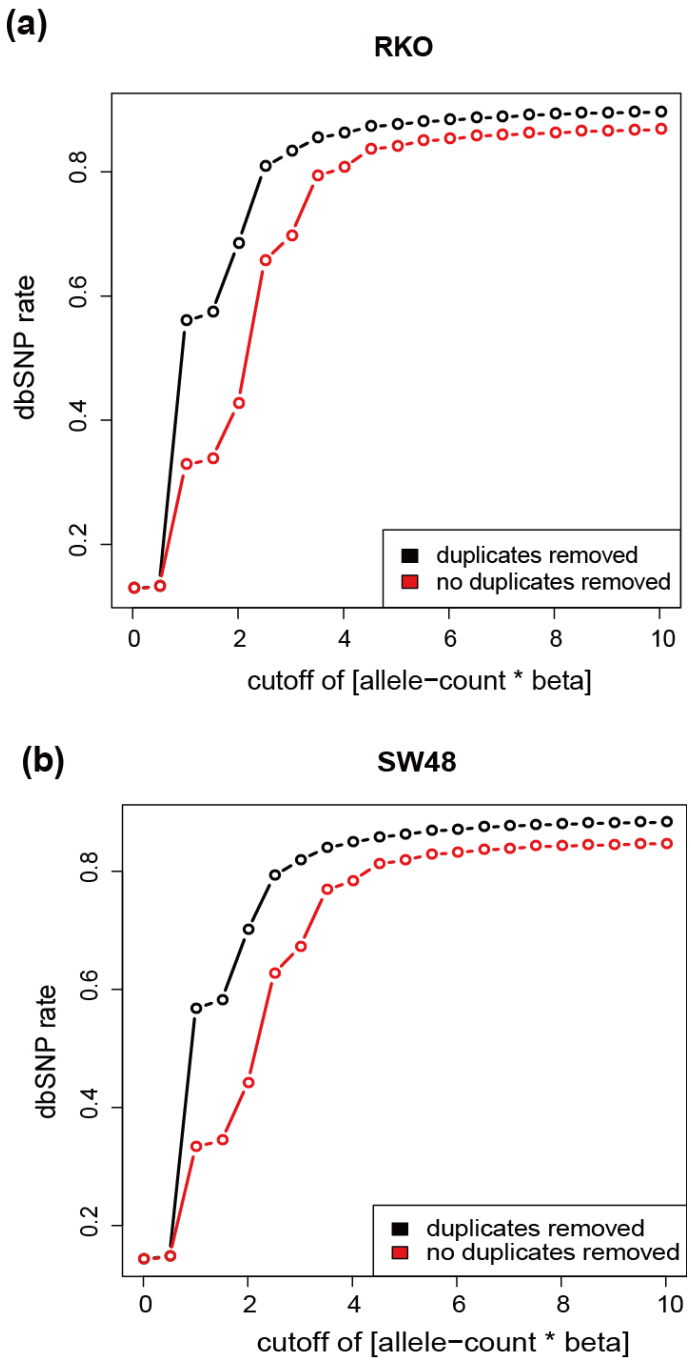


Figure 3.13 Removing the Duplicates increase the dbSNP rate significantly. At various cutoff point of allele-count * β, the dbSNP rates after removing the duplicates become higher.

Table 3.5 Comparison of different tools for calling SNPs in RKO and SW48 cell lines.

		our method	Varscan	DNAnexus
RKO	**total SNP**	95,142	101,088	88,849
	SNPs in dbSNP	76,836	76,560	52,923
	dbSNP rate (%)	80.76	75.74	59.57
	Ti/Tv ratio	2.15	2.22	1.38
	FDR	0.10	0.07	0.52
	($Ti/Tv_{expected}$ = 2.34**)**			
SW48	**total SNP**	113,195	121,494	94,958
	SNPs in dbSNP	92,706	86,014	56,862
	dbSNP rate (%)	81.90	70.80	59.88
	Ti/Tv ratio	2.13	2.07	1.42
	FDR	0.09	0.13	0.49
	($Ti/Tv_{expected}$ = 2.3**)**			

DNAnexus, which is consistent to the dbSNP rates. Note that even the Ti/Tv ratios calculated by Varscan and our method are only ~ 2.2, which is far smaller than the expected Ti/Tv ratio for exome (3.0-3.5). This is because that in the RKO and SW48 datasets a lot of reads are from non-exonic region, reducing the Ti/Tv ratio. The Ti/Tv ratio for the SNPs described in dbSNP is ~2.3, proving that the low Ti/Tv ratio is not caused by inaccurate call. Since ~ 75% of the SNPs are from intronic and intergenic region, we modify $Ti/Tv_{expected}$ to 2.34 according the percentage of exonic SNPs in total SNPs. The FDRs calculated with the modified $Ti/Tv_{expected}$ are consistent with the dbSNP rates. Similar results are obtained for SW48.

4.3.3. Annotation and comparison of SNPs in RKO and SW48

We compared the SNPs of RKO and SW48, and found 32,224 SNPs in common, which is 33.87% and 28.47% of SNPs in RKO and SW48 respectively. Among these common SNPs, 28,326 (87.90%) SNPs are found in dbSNP (**Table 3.6**). With the same method we applied to PC-3, we annotate the SNPs in RKO and SW48 and classified them by their functions, then we found that SNPs from these two datasets distribute in a very similar pattern (**Figure 3.14 (a)**), which is possibly because these two datasets share the same experiment protocol. ~ 50%-60% SNPs are intronic and 20% SNPs are intergenic, which explained the low Ti/Tv ratios. Only about 15% SNPs are from exome. The dbSNP rate varies for the SNPs with different functions (**Figure 3.14 (b)**). Similar to PC-3 data, the dbSNP rate for splicing SNPs are much lower than the others. Exonic SNPs are grouped by their effect to genetic coding (**Table 3.7**). 47 stopgain

Table 3.6 Classification of RKO and SW48 SNPs and their common SNPs.

	PC3			SW48			Common	
	Total SNPs	SNPs in dbSNP	dbSNP rate	Total SNPs	SNPs in dbSNP	dbSNP rate	Total SNPs	SNPs in dbSNP
downstream	705	529	75.04	830	612	73.73	188	152
exonic	11,583	10,471	90.40	10,206	9,315	91.27	4,778	4734
exonic;splicing	159	128	80.50	129	110	85.27	58	57
intergenic	18,490	13,648	73.81	22,731	17,217	75.74	5,429	3887
intronic	54,187	44,871	82.81	68,775	57,747	83.97	18,266	16787
ncRNA_exonic	924	705	76.30	862	674	78.19	362	270
ncRNA_intronic	2,408	1,844	76.58	2,911	2,294	78.80	735	625
ncRNA_splicing	19	4	21.05	14	3	21.43	7	1
ncRNA_UTR3	83	62	74.70	88	71	80.68	32	30
ncRNA_UTR5	10	8	80.00	9	9	100.00	3	3
splicing	921	50	5.43	1,008	47	4.66	505	31
upstream	631	517	81.93	532	472	88.72	189	178
upstream;downstream	51	38	74.51	47	32	68.09	16	12
UTR3	4,161	3,274	78.68	4,317	3,466	80.29	1,353	1274
UTR5	808	686	84.90	736	637	86.55	303	285
UTR5;UTR3	2	1	50.00	0	0		0	0
total	95,142	76,836		113,195	92,706		32,224	28,326

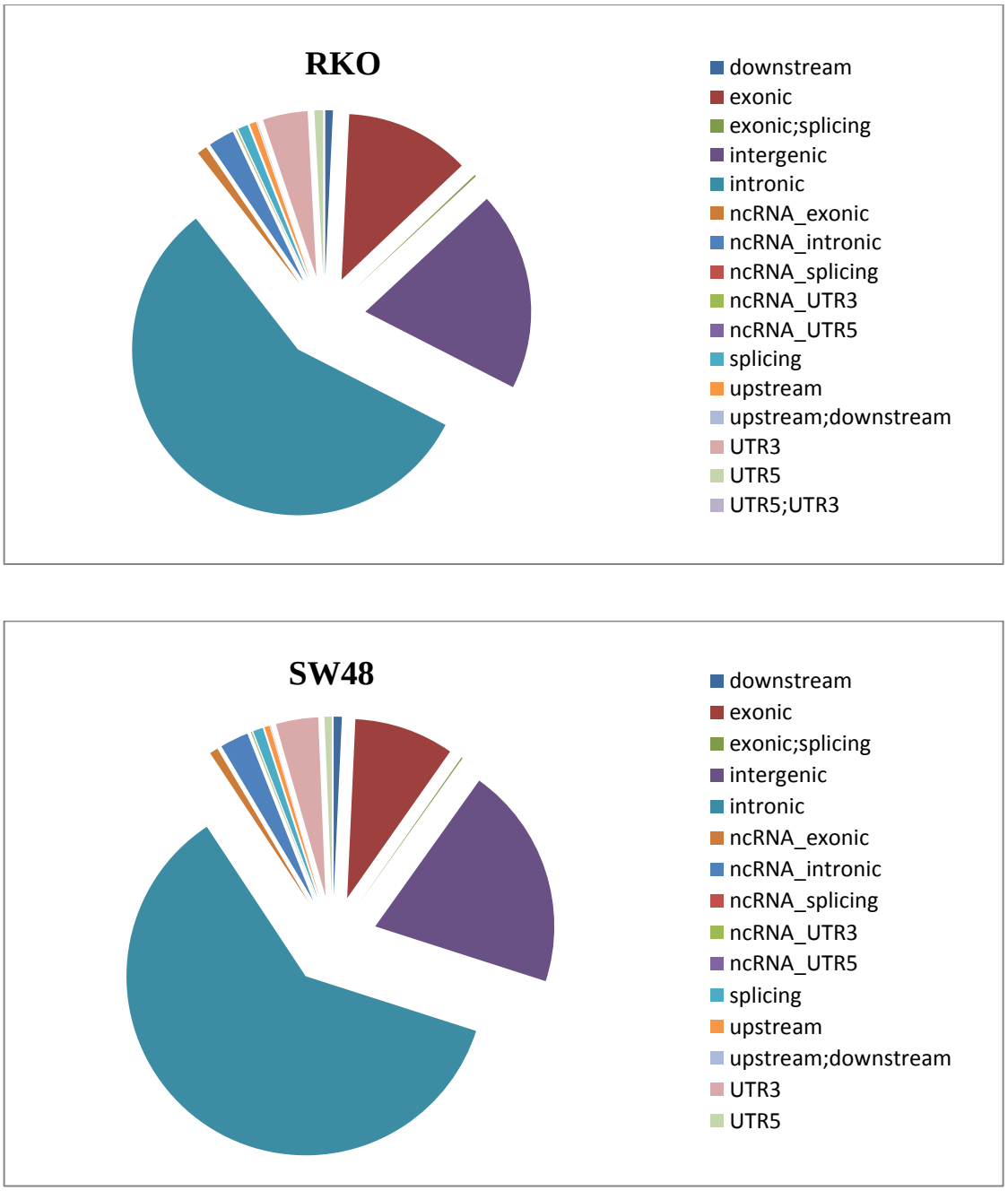
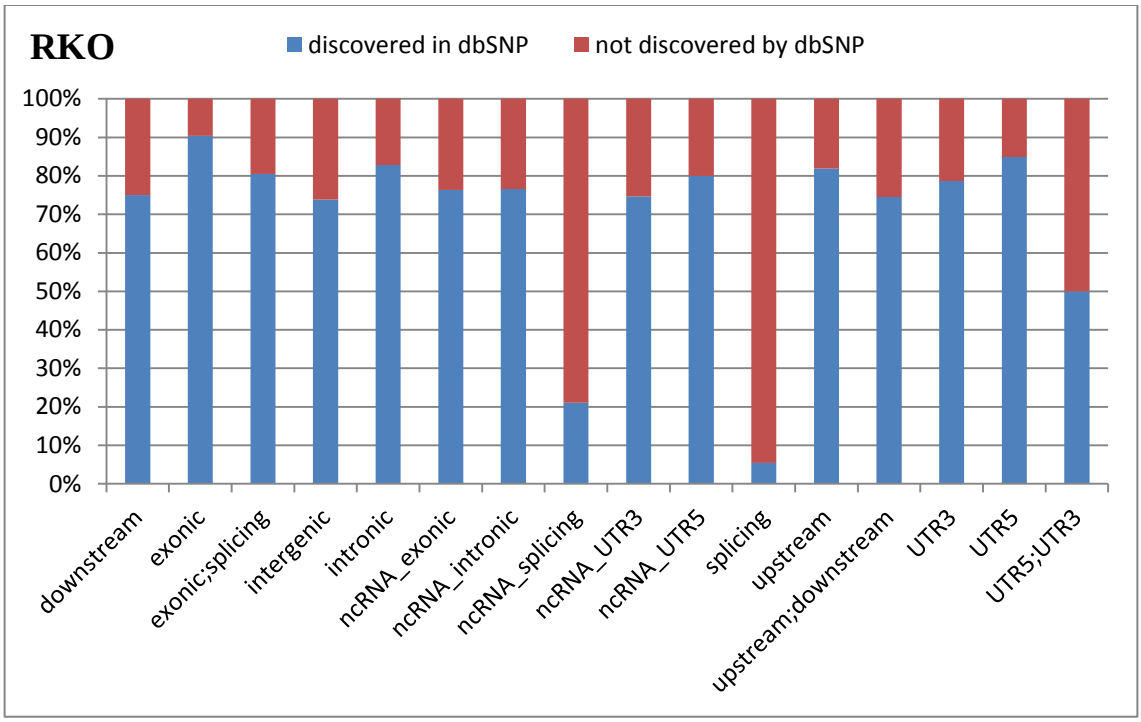


Figure 3.14 (a) Pie charts of RKO and SW48 SNPs for different genome regions.



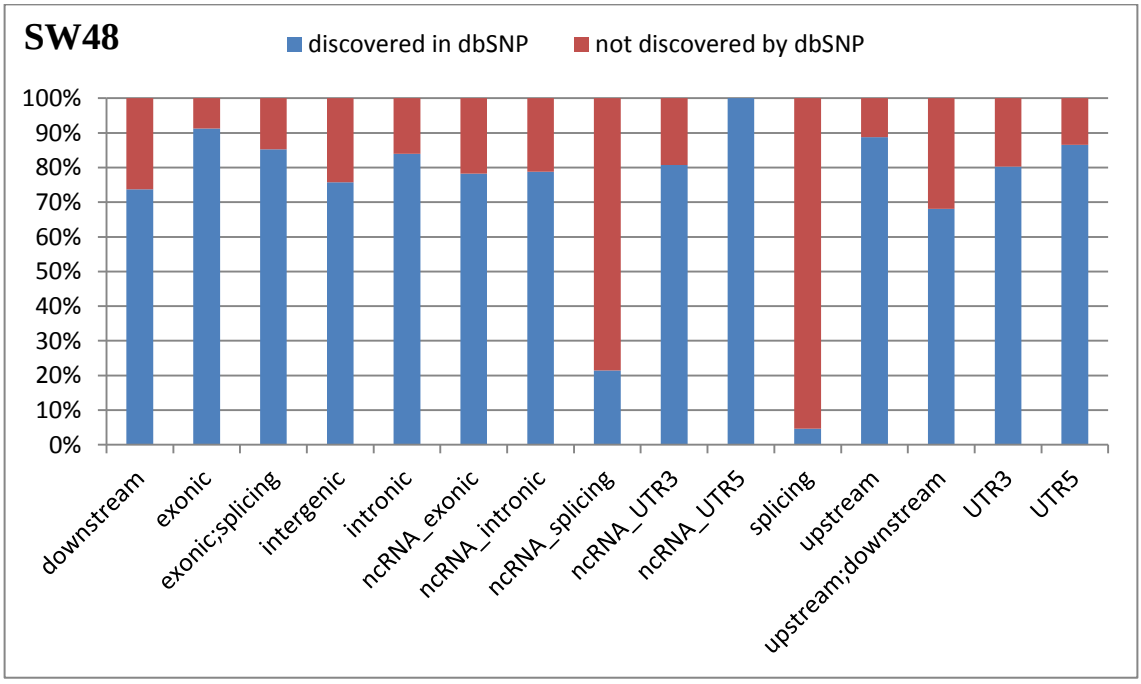


Figure 3.14 (b) Percentage of RKO and SW48 SNPs observed in dbSNP for different regions.

Table 3.7 Classification of exonic SNPs in RKO and SW48 by genetic coding.

	RKO		SW48		common	
	total	**in dbSNP**	**total**	**in dbSNP**	**total**	**in dbSNP**
nonsynonymous	5,563	4,781	4,814	4,195	2,230	2,201
synonymous	6,070	5,738	5,393	5,141	2,566	2,552
stopgain	47	24	54	27	12	12
stoploss	8	8	9	7	2	2
unknown	54	48	65	55	26	24

SNPs are found for RKO, and 54 stopgain SNPs are found for SW48. Among these stopgain SNPs, 12 SNPs are in common. Novel stopgain and stoploss SNPs that are not discovered in dbSNP are listed in **Table S4**.

To identify the potential causal genes for RKO and SW48, we applied the same method used for PC-3 to prioritize the SNPs (**Figure 3.15**). The exonic/splicing nonsynonymous SNPs are filtered with 1000-genome database, ESP6500 database, dbSNP135 database and CG69 database. Furthermore, the SNPs that are believed to be benign by SIFT and PolyPhen are removed. For RKO and SW48, we found 1188 and 1213 SNPs respectively, which correspond to 701 and 652 genes. 292 genes are in common for RKO and SW48.

5. DISCUSSION

The high throughput of NGS provides the possibility to completely study the genome-wide variation, but it also challenges the algorithms for variants detection due to its high sequencing error rates. Most SNP calling methods work well for the high-coverage data, so exome sequencing is the most common strategy for the variants study. Exome is the major disease-causing region but only constitutes about 1% of the human genome, so exome can be sequenced deeply without extra cost. However, exome sequencing is only able to detect the variants in the coding region of genes which control the protein function. It is unable to detect the variants from the remained 99% of the human genome, which can affect the gene regulatory and are also associated with diseases. But sequencing the whole genome deeply is not practical right now due to the

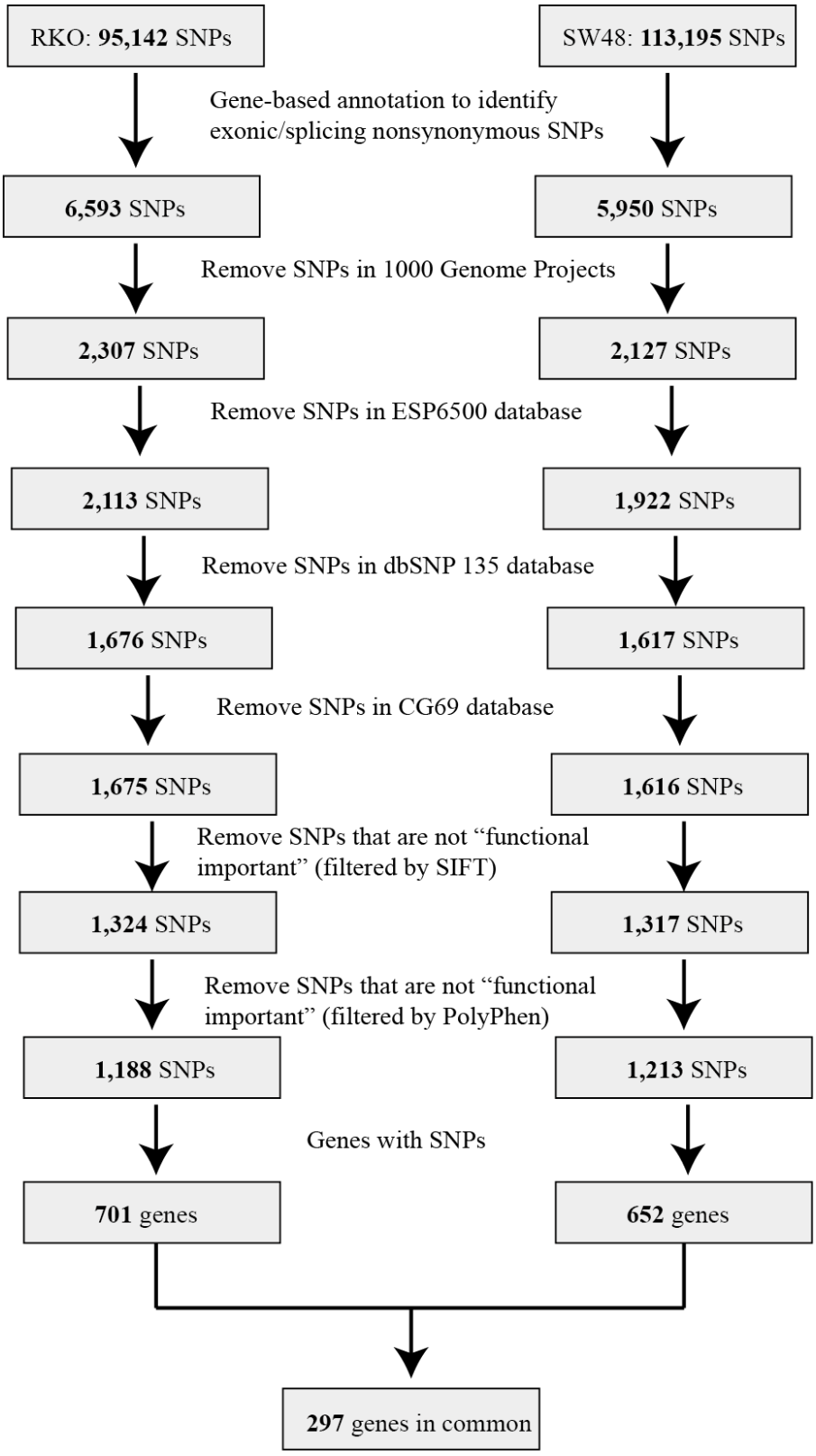


Figure 3.15. Prioritization of causal genes for RKO and SW48.

high cost. Therefore, we report a novel SNP calling algorithm which can achieve high sensitivity and specificity for low coverage data. We have showed that our method perform well on both low coverage ChIP-Seq data and high coverage Exome sequencing data. There are thousands of ChIP-Seq data, RNA-Seq data and some other sequencing data with low or uneven coverage conducted every year, which can also be used for the SNP detection with our method. We expect that by utilizing these data, the catalogue of the human genetic variants will grow fast.

A few verified SNPs have been reported for RKO and SW48. RKO has two verified substitutions: c.1799T>A in BRAF gene and c.3140A>G PIK3CA gene. The first one is shown in our result, but the second was not called because it is the allele with lower frequency and our method only report the primary allele. SW48 also has two verified substitutions: c.98C>A in CTNNB1 gene and c.2155G>A in EGFR gene. Similar to RKO, the first one was identified by our method, but the second one was not called because of its low frequency. The comparison between the verified SNPs and our results shows the reliability of our method. In this article, we only showed the SNPs that are the primary alleles, but our method can also be used to call the alternative allele by reporting the allele with second higher probability. Furthermore, our method is also able to call genotype using the ratio of the highest and second highest probability.

CHAPTER FOUR

Conclusion

In this article, we first present an algorithm that is highly robust and efficient at correcting errors in the NGS data. We demonstrated the effectiveness of MTM on both single-cell data with highly non-uniform coverage and normal data with uniformly high coverage, reflecting that MTM does not rely on the coverage of the sequencing reads. Compared with the previous tools Hammer and Quake, which beat the other tools on non-uniform and uniform data respectively, MTM showed better performance than Hammer and better or similar performance than Quake in terms of the positive predictive value and sensitivity. We also showed the benefits of error correction with MTM on the downstream analysis, such as mapping and SNP detection.

We then present a Bayesian-based approach for SNP calling, which improved the existing methods by having no limitation on the sequencing depth. We successfully applied this approach to identify the SNPs in prostate cancer cell line PC-3 and colon cancer cell lines RKO and SW48, whose mutation status are unknown. Our method outperforms the existing methods - Varscan and DNAnexus - by identifying more SNPs while maintaining higher dbSNP rates, especially for the low coverage PC-3 data. In summary, we identified 107 potential causal genes for PC-3. For RKO and SW48 cell lines, 701 and 652 potential causal genes were identified respectively, and 297 genes are in common. With the ability of piggybacking on the ChIP-Seq, mRNA-Seq and other sequencing data with low or uneven coverage, this approach is expected to have a wide range of applications.

Bibliography

1. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M: **Nucleotide sequence of bacteriophage phi X174 DNA**. *Nature* 1977, **265**(5596):687-695.

2. Swerdlow H, Wu SL, Harke H, Dovichi NJ: **Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette**. *Journal of Chromatography* 1990, **516**(1):61-67.

3. Hunkapiller T, Kaiser RJ, Koop BF, Hood L: **Large-scale and automated DNA sequence determination**. *Science* 1991, **254**(5028):59-67.

4. **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**(7011):931-945.

5. **Pacific Biosciences** [http://www.pacificbiosciences.com]

6. Wetterstrand K: **DNA sequencing cost: data from the NHGRI large-scale genome sequencing program**. *National Human Genome Research Institutue* 2012.

7. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M: **Comparison of next-generation sequencing systems**. *Journal of Biomedicine & Biotechnology* 2012, **2012**:251364.

8. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B: **Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(15):8817-8822.

9. Mitra RD, Church GM: **In situ localized amplification and contact replication of many individual DNA molecules**. *Nucleic Acids Research* 1999, **27**(24):e34.

10. Shendure J, Ji H: **Next-generation DNA sequencing**. *Nat Biotechnol* 2008, **26**(10):1135-1145.

11. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**(7057):376-380.

12. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, Coleman BE, Laptewicz MW, Sannicandro AE, Rhodes MD, Gottimukkala RK, Yang S, Bafna V, Bashir A, MacBride A, Alkan C, Kidd JM, Eichler EE, Reese MG, De La Vega FM, Blanchard AP: **Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding**. *Genome Research* 2009, **19**(9):1527-1541.

13. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ,

Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk

SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ: **Accurate whole human genome sequencing using reversible terminator chemistry**. *Nature* 2008, **456**(7218):53-59.

14. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ: **A large genome center's improvements to the Illumina sequencing system**. *Nat Methods* 2008, **5**(12):1005-1010.

15. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J: **Targeted capture and massively parallel sequencing of 12 human exomes**. *Nature* 2009, **461**(7261):272-276.

16. Ku CS, Naidoo N, Pawitan Y: **Revisiting Mendelian disorders through exome sequencing**. *Hum Genet* 2011, **129**(4):351-370.

17. Morin RD, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh TJ, McDonald H, Varhol R, Jones SJM, Marra MA: **Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing**. *Biotechniques* 2008, **45**(1):81-94.

18. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature Reviews Genetics* 2009, **10**(1):57-63.

19. Maher CA, Kumar-Sinha C, Cao XH, Kalyana-Sundaram S, Han B, Jing XJ, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM: **Transcriptome sequencing to detect gene fusions in cancer**. *Nature* 2009, **458**(7234):97-U99.

20. **Exome sequencing** [http://en.wikipedia.org/wiki/Exome_sequencing]

21. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nat Methods* 2008, **5**(7):621-628.

22. **RNA-Seq** [http://en.wikipedia.org/wiki/RNA_seq]

23. **ChIP-sequencing** [http://en.wikipedia.org/wiki/Chip-Seq]

24. Frazer KA, Murray SS, Schork NJ, Topol EJ: **Human genetic variation and its contribution to complex traits**. *Nature Reviews Genetics* 2009, **10**(4):241-251.

25. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen HC, Agarwala R, McLaren WM, Ritchie GR, Albracht D, Kremitzki M, Rock S, Kotkiewicz H, Kremitzki C, Wollam A, Trani L, Fulton L, Fulton R, Matthews L, Whitehead S, Chow W, Torrance J, Dunn M, Harden G, Threadgold G, Wood J, Collins J, Heath P, Griffiths G, Pelan S, Grafham D, Eichler EE, Weinstock G, Mardis ER, Wilson RK, Howe K, Flicek P, Hubbard T: **Modernizing Reference Genome Assemblies**. *Plos Biol* 2011, **9**(7).

26. Kidd JM, Sampas N, Antonacci F, Graves T, Fulton R, Hayden HS, Alkan C, Malig M, Ventura M, Giannuzzi G, Kallicki J, Anderson P, Tsalenko A, Yamada NA, Tsang P, Kaul R, Wilson RK, Bruhn L, Eichler EE: **Characterization of missing human genome sequences and copy-number polymorphic insertions**. *Nature Methods* 2010, **7**(5):365-U347.

27. Warren RL, Sutton GG, Jones SJM, Holt RA: **Assembling millions of short DNA sequences using SSAKE**. *Bioinformatics* 2007, **23**(4):500-501.

28. Chaisson MJ, Pevzner PA: **Short read fragment assembly of bacterial genomes**. *Genome Research* 2008, **18**(2):324-330.

29. Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J: **De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer**. *Genome Research* 2008, **18**(5):802-809.

30. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB: **ALLPATHS: De novo assembly of whole-genome shotgun microreads**. *Genome Research* 2008, **18**(5):810-820.

31. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs**. *Genome Research* 2008, **18**(5):821-829.

32. Siva N: **1000 Genomes project**. *Nat Biotechnol* 2008, **26**(3):256.

33. **Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species**. *The Journal of Heredity* 2009, **100**(6):659-674.

34. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the National Academy of Sciences of the United States of America* 1977, **74**(12):5463-5467.

35. De Bruijn NG: **A combinatorial problem**. *Proc Koninklijke Nederlandse Akademie v Wetenschappen* 1946, **49**:758-764.

36. Pevzner PA, Tang H, Waterman MS: **An Eulerian path approach to DNA fragment assembly**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(17):9748-9753.

37. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing**. *Nucleic Acids Research* 2008, **36**(16):e105.

38. Cole Trapnell SLS: **How to map billions of short reads onto genomes**. *Nature Biotechnology* 2009, **27**:3.

39. Yang X, Chockalingam SP, Aluru S: **A survey of error-correction methods for next-generation sequencing**. *Brief Bioinform* 2012.

40. Chaisson M, Pevzner P, Tang H: **Fragment assembly with short reads**. *Bioinformatics* 2004, **20**(13):2067-2074.

41. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB: **ALLPATHS: de novo assembly of whole-genome shotgun microreads**. *Genome Research* 2008, **18**(5):810-820.

42. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J: **De novo assembly of human genomes with massively parallel short read sequencing**. *Genome Research* 2010, **20**(2):265-272.

43. Kelley DR, Schatz MC, Salzberg SL: **Quake: quality-aware detection and correction of sequencing errors**. *Genome Biol* 2010, **11**(11):R116.

44. Yang X, Dorman KS, Aluru S: **Reptile: representative tiling for short read error correction**. *Bioinformatics* 2010, **26**(20):2526-2533.

45. Medvedev P, Scott E, Kakaradov B, Pevzner P: **Error correction of high-throughput sequencing datasets with non-uniform coverage**. *Bioinformatics*, **27**(13):i137-141.

46. Shi H, Schmidt B, Liu W, Muller-Wittig W: **A parallel algorithm for error correction in high-throughput short-read data on CUDA-enabled graphics hardware**. *Journal of Computational Biology* 2010, **17**(4):603-615.

47. Zhao X, Palmer LE, Bolanos R, Mircean C, Fasulo D, Wittenberg GM: **EDAR: an efficient error detection and removal algorithm for next generation sequencing data**. *Journal of Computational Biology* 2010, **17**(11):1549-1560.

48. Schroder J, Schroder H, Puglisi SJ, Sinha R, Schmidt B: **SHREC: a short-read error correction method**. *Bioinformatics* 2009, **25**(17):2157-2163.

49. Ilie L, Fazayeli F, Ilie S: **HiTEC: accurate error correction in high-throughput sequencing data**. *Bioinformatics* 2011, **27**(3):295-302.

50. Salmela L: **Correction of sequencing errors in a mixed set of reads**. *Bioinformatics* 2010, **26**(10):1284-1290.

51. Kao WC, Chan AH, Song YS: **ECHO: a reference-free short-read error correction algorithm**. *Genome Research* 2011, **21**(7):1181-1192.

52. Salmela L, Schroder J: **Correcting errors in short reads by multiple alignments**. *Bioinformatics* 2011, **27**(11):1455-1461.

53. Qu W, Hashimoto S, Morishita S: **Efficient frequency-based de novo short-read clustering for error trimming in next-generation sequencing**. *Genome Research* 2009, **19**(7):1309-1315.

54. Wijaya E, Frith MC, Suzuki Y, Horton P: **Recount: expectation maximization based error correction tool for next generation sequencing data**. *Genome Inform* 2009, **23**(1):189-201.

55. Chitsaz H, Yee-Greenbaum JL, Tesler G, Lombardo MJ, Dupont CL, Badger JH, Novotny M, Rusch DB, Fraser LJ, Gormley NA, Schulz-Trieglaff O, Smith GP, Evers DJ, Pevzner PA, Lasken RS: **Efficient de novo assembly of single-cell bacterial genomes from short-read data sets**. *Nat Biotechnol* 2011, **29**(10):915-921.

56. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**(3):R25.

57. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

58. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes**. *Genome Biol* 2004, **5**(2):R12.

59. Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and SNP calling from next-generation sequencing data**. *Nature Reviews Genetics* 2011, **12**(6):443-451.

60. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. *Genome Research* 2008, **18**(11):1851-1858.

61. Li JB, Gao Y, Aach J, Zhang K, Kryukov GV, Xie B, Ahlford A, Yoon JK, Rosenbaum AM, Zaranek AW, LeProust E, Sunyaev SR, Church GM: **Multiplex padlock targeted sequencing reveals human hypermutable CpG variations**. *Genome Research* 2009, **19**(9):1606-1615.

62. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment**. *Bioinformatics* 2009, **25**(15):1966-1967.

63. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K: **SNP detection for massively parallel whole-genome resequencing**. *Genome Research* 2009, **19**(6):1124-1132.

64. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ: **A framework for variation discovery and genotyping using next-generation DNA sequencing data**. *Nature Genetics* 2011, **43**(5):491-498.

65. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L: **VarScan: variant detection in massively parallel sequencing of individual and pooled samples**. *Bioinformatics* 2009, **25**(17):2283-2285.

66. Erlich Y, Mitra PP, delaBastide M, McCombie WR, Hannon GJ: **Alta-Cyclic: a self-optimizing base caller for next-generation sequencing**. *Nat Methods* 2008, **5**(8):679-682.

67. Rougemont J, Amzallag A, Iseli C, Farinelli L, Xenarios I, Naef F: **Probabilistic base calling of Solexa sequencing data**. *BMC Bioinformatics* 2008, **9**:431.

68. Kaighn ME, Lechner JF, Narayan KS, Jones LW: **Prostate carcinoma: tissue culture cell lines**. *Natl Cancer Inst Monogr* 1978(49):17-21.

69. Kaighn ME, Narayan KS, Ohnuki Y, Lechner JF, Jones LW: **Establishment and characterization of a human prostatic carcinoma cell line (PC-3)**. *Investigative Urology* 1979, **17**(1):16-23.

70. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**(6):841-842.

71. **SNP Call Set Properties**

[http://genome.sph.umich.edu/wiki/SNP_Call_Set_Properties]

72. **The Current State of dbSNP** [http://massgenomics.org/2012/01/the-current-state-of-dbsnp.html]

73. Wakeley J: **The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance**. *Trends Ecol Evol* 1996, **11**(4):158-163.

74. Wakeley J: **Substitution-Rate Variation among Sites and the Estimation of Transition Bias**. *Molecular Biology and Evolution* 1994, **11**(3):436-442.

75. Curtis SE, Clegg MT: **Molecular evolution of chloroplast DNA sequences**. *Mol Biol Evol* 1984, **1**(4):291-301.

76. Gojobori T, Li WH, Graur D: **Patterns of nucleotide substitution in pseudogenes and functional genes**. *J Mol Evol* 1982, **18**(5):360-369.

77. Brown WM, Prager EM, Wang A, Wilson AC: **Mitochondrial DNA sequences of primates: tempo and mode of evolution**. *J Mol Evol* 1982, **18**(4):225-239.

78. Freudenberg-Hua Y, Freudenberg J, Kluck N, Cichon S, Propping P, Nothen MM: **Single nucleotide variation analysis in 65 candidate genes for CNS disorders in a representative sample of the European population**. *Genome Research* 2003, **13**(10):2271-2276.

79. Ebersberger I, Metzler D, Schwarz C, Paabo S: **Genomewide comparison of DNA sequences between humans and chimpanzees**. *Am J Hum Genet* 2002, **70**(6):1490-1497.

80. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. *Nucleic Acids Research* 2010, **38**(16):e164.

81. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloğlu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP: **Genetic diagnosis by whole exome capture and massively parallel DNA sequencing**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(45):19096-19101.

82. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ: **Exome sequencing identifies the cause of a mendelian disorder**. *Nature Genetics* 2010, **42**(1):30-35.

83. Ng PC, Henikoff S: **SIFT: predicting amino acid changes that affect protein function**. *Nucleic Acids Research* 2003, **31**(13):3812-3814.

84. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations**. *Nat Methods* 2010, **7**(4):248-249.

Supplementary

Table S1. Data source of PC3 used in SNPs identification.

Data ID	Lane #	Cell line	Experiment
090218_HWUSI-EAS182R_0001_30KPG	1	PC3	LMN
	2	PC3	Pol2
	3	PC3	H3K4me3
	4	PC3	H3K9me2
	5	PC3	H3K27me3
	6	PC3	AGO2
	7	PC3	Total_H3
090220_HWUSI-EAS182R_0002_30KR3	1	PC3	LMN
	2	PC3	Pol2
	3	PC3	H3K4me3
	4	PC3	H3K9me2
	5	PC3	H3K27me3
	6	PC3	AGO2
	7	PC3	H3
090512_HWUSI-EAS230-R_0003_30NYJ	1	PC3	LMN
	2	PC3	Pol2
	3	PC3	H3K4me3
	4	PC3	H3K9me2
	5	PC3	H3K27me3
	6	PC3	AGO2
	7	PC3	Total_H3
090526_HWUSI-EAS230-R_0006_30P04	1	PC3	H3K9me2
	2	PC3	H3K9me2
	3	PC3	H3K9me2
	4	PC3	H3K9me3
	5	PC3	H3K9me3
	6	PC3	H3K9me3
090327_HWUSI-EAS182R_0006_30M3H	3	PC3	input
	4	PC3	H3
	5	PC3	H3K4me3
	6	PC3	H3K27me3
	7	PC3	input
090327_HWUSI-EAS182R_0006_30MKR	1	PC3	input
	2	PC3	H3
	3	PC3	H3K4me3
	4	PC3	H3K27me3
	5	PC3	input
090526_HWUSI-EAS230-R_0006_30P04	7	PC3	H3K27me3

090507_HWUSI-EAS230-R_0002_30NWT	6	PC3	H3
	7	PC3	H3K4me3

Table S2. Noval stopgain and stoploss SNPs for PC-3

function	chromosome	position	ref	SNP	Gene information
stopgain	chr3	160155983	A	T	TRIM59:NM_173084:exon3:c.T989A:p.L330X,
stopgain	chr12	56647988	C	A	ANKRD52:NM_173595:exon8:c.G769T:p.E257X,
stopgain	chr16	1822802	G	A	MRPS34:NM_023936:exon1:c.C319T:p.Q107X,
stopgain	chr19	50916763	C	A	POLD1:NM_001256849:exon18:c.C2235A:p.Y745X,POLD1:NM_002691:exon18:c.C2235A:p.Y745X,
stoploss	chr19	35633644	T	G	FXYD1:NM_005031:exon7:c.T277G:p.X93E,FXYD1:NM_021902:exon7:c.T277G:p.X93E,

Table S3. Potential causal genes identified by our method for PC-3

Gene Symbol	SNPs number	Entrez Gene Name	Location	Type(s)
CTBP2	3	C-terminal binding protein 2	Nucleus	transcription regulator
ZNF717	2	zinc finger protein 717	Nucleus	other
NOXO1	2	NADPH oxidase organizer 1	Plasma Membrane	other
ALG9	2	asparagine-linked glycosylation 9, alpha-1,2-mannosyltransferase homolog (S. cerevisiae)	Cytoplasm	enzyme
KIR2DL3	2	killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 3	Plasma Membrane	other
RARRES2	2	retinoic acid receptor responder (tazarotene induced) 2	Plasma Membrane	transmembrane receptor
CHD8	2	chromodomain helicase DNA binding protein 8	Nucleus	enzyme
SNX18	2	sorting nexin 18	Cytoplasm	transporter
UNC80	2	unc-80 homolog (C. elegans)	unknown	other
CCDC57	2	coiled-coil domain containing 57	unknown	other
DUSP28	2	dual specificity phosphatase 28	unknown	enzyme
PTPRM	2	protein tyrosine phosphatase, receptor type, M	Plasma Membrane	phosphatase
LLGL1	2	lethal giant larvae homolog 1 (Drosophila)	Cytoplasm	other
FAM174B	2	family with sequence similarity 174, member B	unknown	other
VEPH1	1	ventricular zone expressed PH domain homolog 1 (zebrafish)	Nucleus	other
MMS19	1	MMS19 nucleotide excision repair homolog (S. cerevisiae)	Nucleus	transcription regulator
FREM3	1	FRAS1 related extracellular matrix 3	Extracellular Space	other
IFI35	1	interferon-induced protein 35	Nucleus	other
PLD4	1	phospholipase D family, member 4	Extracellular Space	enzyme
URB1	1	URB1 ribosome biogenesis 1 homolog (S. cerevisiae)	Nucleus	other
GATA2	1	GATA binding protein 2	Nucleus	transcription regulator
DAPP1	1	dual adaptor of phosphotyrosine and 3-phosphoinositides	Cytoplasm	other
DSE	1	dermatan sulfate epimerase	Cytoplasm	enzyme
RUNX3	1	runt-related transcription factor 3	Nucleus	transcription regulator
EYS	1	eyes shut homolog (Drosophila)	unknown	other
THAP3	1	THAP domain containing, apoptosis associated protein 3	unknown	other
RBMXL3	1	RNA binding motif protein, X-linked-like 3	unknown	other
NTN3	1	netrin 3	Extracellular Space	other

FEM1A	1	fem-1 homolog a (C. elegans)	Nucleus	transcription regulator
RAP2A	1	RAP2A, member of RAS oncogene family	Plasma Membrane	enzyme
DMRT1	1	doublesex and mab-3 related transcription factor 1	Nucleus	transcription regulator
TMC2	1	transmembrane channel-like 2	Plasma Membrane	other
OBSCN	1	obscurin, cytoskeletal calmodulin and titin-interacting RhoGEF	Cytoplasm	kinase
EIF4G2	1	eukaryotic translation initiation factor 4 gamma, 2	Cytoplasm	translation regulator
AEN	1	apoptosis enhancing nuclease	Nucleus	enzyme
GFM2	1	G elongation factor, mitochondrial 2	Cytoplasm	translation regulator
COPB2	1	coatomer protein complex, subunit beta 2 (beta prime)	Cytoplasm	transporter
OR4D9	1	olfactory receptor, family 4, subfamily D, member 9	Plasma Membrane	G-protein coupled receptor
KIF5C	1	kinesin family member 5C	Cytoplasm	other
CDC27	1	cell division cycle 27 homolog (S. cerevisiae)	Nucleus	other
SGPP2	1	sphingosine-1-phosphate phosphatase 2	Cytoplasm	phosphatase
CTBP1	1	C-terminal binding protein 1	Nucleus	enzyme
CUL9	1	cullin 9	Cytoplasm	other
UTP18	1	UTP18 small subunit (SSU) processome component homolog (yeast)	Nucleus	other
C17orf100	1	chromosome 17 open reading frame 100	unknown	other
DUSP5	1	dual specificity phosphatase 5	Nucleus	phosphatase
EPPK1	1	epiplakin 1	Cytoplasm	other
RHPN1	1	rhophilin, Rho GTPase binding protein 1	Cytoplasm	other
MLL3	1	myeloid/lymphoid or mixed-lineage leukemia 3	Nucleus	transcription regulator
AP2A1	1	adaptor-related protein complex 2, alpha 1 subunit	Cytoplasm	transporter
LRRIQ1	1	leucine-rich repeats and IQ motif containing 1	unknown	other
CATSPER 1	1	cation channel, sperm associated 1	Plasma Membrane	ion channel
PRSS56	1	protease, serine, 56	unknown	other
ODZ2	1	odz, odd Oz/ten-m homolog 2 (Drosophila)	Plasma Membrane	other
MYH2	1	myosin, heavy chain 2, skeletal muscle, adult	Cytoplasm	enzyme
SNED1	1	sushi, nidogen and EGF-like domains 1	Plasma Membrane	other
G6PD	1	glucose-6-phosphate dehydrogenase	Cytoplasm	enzyme
ONECUT 3	1	one cut homeobox 3	Nucleus	transcription regulator
KLC1	1	kinesin light chain 1	Cytoplasm	other
NLK	1	nemo-like kinase	Nucleus	kinase

MUC2	1	mucin 2, oligomeric mucus/gel-forming	unknown	other
C11orf9	1	chromosome 11 open reading frame 9	Nucleus	transcription regulator
C17orf105	1	chromosome 17 open reading frame 105	unknown	other
SERINC4	1	serine incorporator 4	unknown	other
PPP4C	1	protein phosphatase 4, catalytic subunit	Cytoplasm	phosphatase
C1orf172	1	chromosome 1 open reading frame 172	unknown	other
ABCA6	1	ATP-binding cassette, sub-family A (ABC1), member 6	Plasma Membrane	transporter
FBLN2	1	fibulin 2	Extracellular Space	other
CLDN3	1	claudin 3	Plasma Membrane	transmembrane receptor
ERCC3	1	excision repair cross-complementing rodent repair deficiency, complementation group 3	Nucleus	enzyme
TTLL3	1	tubulin tyrosine ligase-like family, member 3	Extracellular Space	enzyme
KCNH2	1	potassium voltage-gated channel, subfamily H (eag-related), member 2	Plasma Membrane	ion channel
MVP	1	major vault protein	Nucleus	other
MUC6	1	mucin 6, oligomeric mucus/gel-forming	Extracellular Space	other
FAM179A	1	family with sequence similarity 179, member A	unknown	other
GRID1	1	glutamate receptor, ionotropic, delta 1	Plasma Membrane	ion channel
LHX3	1	LIM homeobox 3	Nucleus	transcription regulator
VWF	1	von Willebrand factor	Extracellular Space	other
CTDSP2	1	CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase 2	Nucleus	phosphatase
HK3	1	hexokinase 3 (white cell)	Cytoplasm	kinase
PELP1	1	proline, glutamate and leucine rich protein 1	Nucleus	other
PREX2	1	phosphatidylinositol-3,4,5-trisphosphate-dependent Rac exchange factor 2	Cytoplasm	other
ISYNA1	1	inositol-3-phosphate synthase 1	unknown	enzyme
KCNG2	1	potassium voltage-gated channel, subfamily G, member 2	Plasma Membrane	ion channel
PCNXL2	1	pecanex-like 2 (Drosophila)	unknown	other
HOXD10	1	homeobox D10	Nucleus	transcription regulator
DUX4	1	double homeobox 4	Nucleus	transcription regulator
FSCN1	1	fascin homolog 1, actin-bundling protein (Strongylocentrotus purpuratus)	Cytoplasm	other
ELFN1	1	extracellular leucine-rich repeat and fibronectin type III domain containing 1	unknown	other
ZNF649	1	zinc finger protein 649	Nucleus	other

RYR1	1	ryanodine receptor 1 (skeletal)	Cytoplasm	ion channel
TMEM59L	1	transmembrane protein 59-like	Cytoplasm	other
ABCF3	1	ATP-binding cassette, sub-family F (GCN20), member 3	unknown	transporter
MUC16	1	mucin 16, cell surface associated	unknown	other
PPM1N	1	protein phosphatase, Mg2+/Mn2+ dependent, 1N (putative)	Extracellular Space	other
C7	1	complement component 7	Extracellular Space	other
ANKRD52	1	ankyrin repeat domain 52	Nucleus	transcription regulator
FOXA1	1	forkhead box A1	Nucleus	transcription regulator
WDR34	1	WD repeat domain 34	Cytoplasm	other
FAM59B	1	family with sequence similarity 59, member B	unknown	other
TPST2	1	tyrosylprotein sulfotransferase 2	Cytoplasm	enzyme
ISLR2	1	immunoglobulin superfamily containing leucine-rich repeat 2	Plasma Membrane	other
FAT1	1	FAT tumor suppressor homolog 1 (Drosophila)	Plasma Membrane	other
SAMD11	1	sterile alpha motif domain containing 11	Nucleus	other
SSPO	1	SCO-spondin homolog (Bos taurus)	Cytoplasm	other
KRTAP7	1	unkonwn	unknown	unknown
HLA	1	unkonwn	unknown	unknown

Table S4. Noval stopgain and stoploss SNPs in RKO and SW48

(a) Noval stopgain and stoploss SNPs in RKO

function	chromosome	position	reference	SNP	Gene information
stopgain	chr1	152185725	G	A	HRNR:NM_001009931:exon3:c.C8380T:p.Q2794X,
stopgain	chr2	85255047	C	T	KCMF1:NM_020122:exon2:c.C52T:p.R18X,
stopgain	chr2	220160985	C	T	PTPRN:NM_001199764:exon18:c.G2201A:p.W734X,PTPRN:NM_001199763:exon17:c.G2384A:p.W795X,PTPRN:NM_002846:exon18:c.G2471A:p.W824X,
stopgain	chr3	40211492	G	T	MYRIP:NM_015460:exon8:c.G781T:p.G261X,
stopgain	chr3	98109948	G	T	OR5K3:NM_001005516:exon1:c.G439T:p.G147X,
stopgain	chr4	56225592	G	T	SRD5A3:NM_024592:exon2:c.G301T:p.G101X,
stopgain	chr5	16671018	G	A	MYO10:NM_012334:exon39:c.C5500T:p.R1834X,
stopgain	chr5	140579957	G	T	PCDHB11:NM_018931:exon1:c.G610T:p.E204X,
stopgain	chr6	52696720	G	A	GSTA5:NM_153699:exon7:c.C595T:p.Q199X,
stopgain	chr6	152457756	G	T	SYNE1:NM_033071:exon141:c.C25512A:p.C8504X,SYNE1:NM_182961:exon141:c.C25656A:p.C8552X,
stopgain	chr10	5773040	C	T	FAM208B:NM_017782:exon11:c.C1078T:p.Q360X,
stopgain	chr12	85279758	G	A	SLC6A15:NM_018057:exon3:c.C379T:p.R127X,SLC6A15:NM_182767:exon3:c.C379T:p.R127X,SLC6A15:NM_001146335:exon2:c.C58T:p.R20X,
stopgain	chr12	112330854	G	T	MAPKAPK5:NM_139078:exon14:c.G1411T:p.E471X,MAPKAPK5:NM_003668:exon14:c.G1405T:p.E469X,
stopgain	chr13	47361163	C	T	ESD:NM_001984:exon4:c.G150A:p.W50X,
stopgain	chr14	73491163	G	T	ZFYVE1:NM_021260:exon2:c.C54A:p.C18X,
stopgain	chr16	50745326	C	T	NOD2:NM_022162:exon4:c.C1504T:p.Q502X,
stopgain	chr17	7186639	C	T	SLC2A4:NM_001042:exon2:c.C109T:p.Q37X,
stopgain	chr17	27049838	A	T	RPL23A:NM_000984:exon3:c.A307T:p.K103X,
stopgain	chr17	62856696	C	A	LRRC37A3:NM_199340:exon11:c.G3568T:p.E1190X,

stopgain	chr17	78172484	C	T	CARD14:NM_052819:exon10:c.C1234T:p.R412X,
stopgain	chr19	45575836	G	A	ZNF296:NM_145288:exon3:c.C451T:p.R151X,
stopgain	chr20	10279932	C	T	SNAP25:NM_130811:exon7:c.C424T:p.R142X,SNAP25:NM_003081:exon7:c.C424T:p.R142X,
stopgain	chr22	31331037	G	A	MORC2:NM_014941:exon20:c.C1738T:p.R580X,

(a) Noval stopgain and stoploss SNPs in SW48

function	chromosome	position	reference	SNP	Gene information
stopgain	chr1	15888764	C	T	DNAJC16:NM_015291:exon9:c.C1282T:p.Q428X,
stopgain	chr1	85561701	C	T	WDR63:NM_145172:exon11:c.C1261T:p.Q421X,
stopgain	chr1	108742652	C	A	SLC25A24:NM_013386:exon1:c.G109T:p.G37X,
stopgain	chr1	151751716	T	A	TDRKH:NM_006862:exon5:c.A424T:p.R142X,TDRKH:NM_001083963:exon5:c.A424T:p.R142X,TDRKH:NM_001083965:exon5:c.A424T:p.R142X,
stopgain	chr1	157516869	C	T	FCRL5:NM_031281:exon3:c.G171A:p.W57X,FCRL5:NM_001195388:exon3:c.G171A:p.W57X,
stopgain	chr2	54885067	C	T	SPTBN1:NM_178313:exon29:c.C6088T:p.R2030X,SPTBN1:NM_003128:exon30:c.C6127T:p.R2043X,
stopgain	chr2	149543893	A	T	EPC2:NM_015630:exon14:c.A2371T:p.R791X,
stopgain	chr2	198949979	C	T	PLCL1:NM_006226:exon2:c.C1738T:p.R580X,
stopgain	chr2	236945333	C	T	AGAP1:NM_001037131:exon14:c.C1774T:p.Q592X,AGAP1:NM_014914:exon13:c.C1615T:p.Q539X,
stopgain	chr3	98304503	C	T	CPOX:NM_000097:exon5:c.G954A:p.W318X,
stopgain	chr3	113374426	C	A	KIAA2018:NM_001009899:exon7:c.G6103T:p.G2035X,
stopgain	chr5	36036015	G	A	UGT3A2:NM_001168316:exon6:c.C1255T:p.Q419X,UGT3A2:NM_174914:exon7:c.C1357T:p.Q453X,
stopgain	chr5	98129453	T	A	RGMB:NM_001012761:exon5:c.T1433A:p.L478X,
stopgain	chr5	112901596	C	T	YTHDC2:NM_022828:exon21:c.C2722T:p.Q908X,
stopgain	chr6	152476161	G	A	SYNE1:NM_033071:exon132:

					c.C23782T:p.R7928X,SYNE1:NM_182961:exon133:c.C23995T:p.R7999X,
stopgain	chr9	135601233	G	A	AK8:NM_152572:exon13:c.C1282T:p.Q428X,
stopgain	chr11	124765398	G	A	ROBO4:NM_019055:exon6:c.C991T:p.R331X,
stopgain	chr12	56295916	C	A	WIBG:NM_001143853:exon3:c.G352T:p.E118X,WIBG:NM_032345:exon3:c.G355T:p.E119X,
stopgain	chr15	45003808	C	T	B2M:NM_004048:exon1:c.C64T:p.Q22X,
stopgain	chr15	89862496	G	A	POLG:NM_001126131:exon19:c.C3067T:p.Q1023X,POLG:NM_002693:exon19:c.C3067T:p.Q1023X,
stopgain	chr16	30581722	C	A	ZNF688:NM_145271:exon3:c.G346T:p.E116X,ZNF688:NM_001024683:exon3:c.G304T:p.E102X,
stopgain	chr17	7673613	C	T	DNAH2:NM_020877:exon24:c.C3985T:p.Q1329X,
stopgain	chr19	34941217	G	A	UBA2:NM_005499:exon9:c.G819A:p.W273X,
stopgain	chr19	39219995	C	T	ACTN4:NM_004924:exon21:c.C2659T:p.Q887X,
stopgain	chr19	49949888	G	A	PIH1D1:NM_017916:exon8:c.C751T:p.Q251X,
stopgain	chrX	100417941	C	A	CENPI:NM_006733:exon21:c.C2256A:p.C752X,
stopgain	chrX	117762191	G	T	DOCK11:NM_144658:exon34:c.G3730T:p.E1244X,
stoploss	chr1	203137787	T	C	MYBPH:NM_004997:exon10:c.A1434G:p.X478W,

VITA

Xiaofeng Zheng was born on April 21st, 1980 in Benxi, China, the daughter of Zhenbo Zheng and Yahua Bi. She got her B.S Degree in Biological Science in 2003 and M.S Degree in Biochemistry & Molecular Biology in 2006 from the School of Life Sciences at Nanjing University. In August of 2006 she entered University of Texas Health Science Center at Houston Graduate School of Biomedical Sciences.

Permanent address:

Chunming Road, Building 476

Benxi

Liaoning, China, 117000